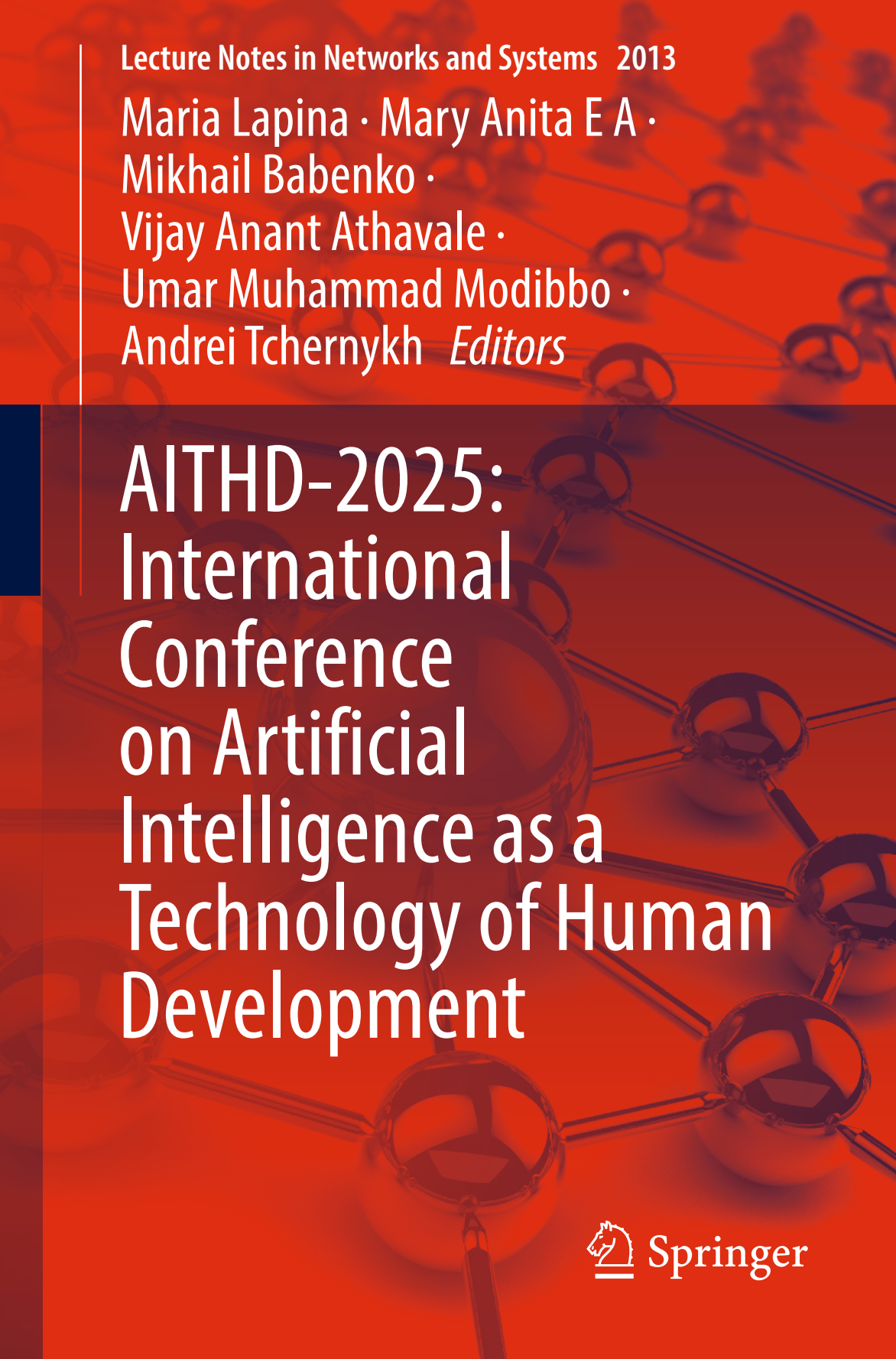
Lecture Notes in Networks and Systems 2013

Maria Lapina · Mary Anita E A ·
Mikhail Babenko ·
Vijay Anant Athavale ·
Umar Muhammad Modibbo ·
Andrei Tchernykh *Editors*

AITHD-2025: International Conference on Artificial Intelligence as a Technology of Human Development



Springer

Lecture Notes in Networks and Systems 2013

Series Editor

Janusz Kacprzyk, *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Advisory Editors

Fernando Gomide, *Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil*
Okyay Kaynak, *Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Türkiye*
Derong Liu, *Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA*
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Witold Pedrycz, *Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada*
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
Marios M. Polycarpou, *Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus*
Imre J. Rudas, *Óbuda University, Budapest, Hungary*
Jun Wang, *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Maria Lapina · Mary Anita E A ·
Mikhail Babenko · Vijay Anant Athavale ·
Umar Muhammad Modibbo · Andrei Tchernykh
Editors

AITHD-2025: International Conference on Artificial Intelligence as a Technology of Human Development

Springer

Editors
Maria Lapina
North-Caucasus Federal University
Stavropol, Russia

Mikhail Babenko
North-Caucasus Federal University
Stavropol, Russia

Umar Muhammad Modibbo
Modibbo Adama University
Yola, Nigeria

Mary Anita E A
Christ University
Bengaluru, Karnataka, India

Vijay Anant Athavale
KLS Gogte Institute of Technology
Belagavi, Karnataka, India

Andrei Tchernykh
CICESE Research Center
Ensenada, Mexico

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-032-27833-3 ISBN 978-3-032-27834-0 (eBook)
https://doi.org/10.1007/978-3-032-27834-0

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2027

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Acknowledgements

We would like to thank all the authors for their submissions, presentation of the results of scientific research, and our Program Committee members and additional reviewers for their precious work.

About This Book

This book is based on the best papers accepted for presentation during the AITHD-2025: International Conference on Artificial intelligence as a Technologies of Human Development on September 17–19, 2025, in Vernadsky Federal University and North Caucasus Federal University, Russia, and online.

The book presents original research at the intersection of artificial intelligence and human development, exploring how machine learning technologies are reshaping industries and enhancing human capabilities. The collection covers the development of neural network classifiers for various tasks, including transport analysis, emotion recognition, and land use classification, as well as approaches to analyzing user preferences and literary texts using machine learning technologies.

Significant attention is given to cybersecurity problems: anomaly detection in information systems, comparative analysis of deep learning methods for intrusion detection in IoT networks, phishing website detection, and malware detection. Hybrid cryptographic approaches combining post-quantum and classical methods are considered, along with cyber resilience assessment of critical information infrastructure facilities.

The intersection of AI with human health becomes a central theme. Research on brain age estimation, driver drowsiness detection systems based on physiological data, speaker recognition under noisy conditions, and mental workload assessment in multiple sclerosis patients is presented. Work on multimodal sentiment analysis, automated diagnosis of mental disorders based on speech biomarker analysis, and epileptic seizure detection using neural networks is included.

Applications in the food industry and agriculture demonstrate AI's role in optimizing winemaking processes, predicting antioxidant activity of wine materials, and IoT-enabled precision farming for sustainable agriculture.

The transformation of education through AI is examined through research on adaptive foreign language learning systems, tools for academic vocabulary learning, and teacher training for AI integration in professional activities.

Fundamental research includes bias problems in recommendation systems, fake news detection methods, differential privacy mechanisms, image restoration using diffusion models, and optimization of quantum key distribution protocols.

Media and social impact studies cover advertising effectiveness analysis in tourism, the convergence of AI and telecommunication systems, machine vision market analysis, and automated hero image monitoring systems in social networks. Issues of cyberbullying and online safety are also addressed.

The book is intended for specialists in artificial intelligence applications, machine learning engineering, information security, data science, educational technology, and researchers at the intersection of technological innovation and human development.

Contents

Intelligent, Lightweight and Secure Authentication Scheme for Smart Home 1
Irfan Alam, Zubair Ashraf, Mohammad Shahid, and Maria Lapina

Comparative Analysis of the Effectiveness of Various Machine Learning Models in Content Generation Tasks 12
Yulia Andrusenko, Ivan Golovin, Elena Dombrovskaya, and Mary Anita E A

Deepfakes as a Cyberattack Tool: New Vectors of Social Engineering 24
Yulia Andrusenko, Dmitry Lukyanov, Ivan Golovin, Elena Maksimova, and Maria Ndapewa Ntinda

Development of a Classifier of Motor Transport Based on Neural Networks 32
Ivan Azarov, Elmira Azarova, Alexey Dmitrienko, Alexey Kozhevnikov, and Mohammed Abdul Qadeer

Research and Development of a Software Module for Analyzing the Preferences of Online Store Users 44
Ivan Azarov, Elmira Azarova, Marina Tokmakova, Magomed Abdurazakov, and Umar Muhammad Modibbo

Development of a Classifier of Human Emotions Based on Neural Networks . . . 58
Ivan Azarov, Elmira Azarova, Marina Tokmakova, Shamil Barakaev, and Antony Shibu Paul

Approaches to the Analysis of Literary Texts Using Machine Learning Technologies 72
Irina Babenko

Analysing Deep Learning Models for Detecting Anomalies in Video Surveillance Data 83
Vinodha D, Mary Anita E A, Jenefa J, and Vladimir Lapin

Direct Response Advertising for Small Businesses in the Tourism Industry: An Analysis of Effectiveness (Based on the Promotion of a Custom Tour to the Caucasus) 92
Anton Degtyarev, Elena Yezhova, Alisa Kushschenko, and Prakasha G S

Comparative Analysis of Deep Learning and Ensemble Methods for Intrusion Detection in IoT Networks 103
Melsa Magdalene Fernandes, Sharlene Anna Pereira, Mary Anita E A, and Vinodha D

Actualization of the "Hero Image" as a Consequence of the Convergence of Artificial Intelligence and Telecommunication Systems 112
Oleg Gabrielyan, Yelizaveta Vitulyova, and Ibragim Suleimenov

Optimization of Winemaking Processes Using Artificial Intelligence Methods 123
Nadezhda Gallini, Anatoliy Kazak, Yuriy Grishin, Marina Rudenko, and Victor Gallini

Application of Artificial Intelligence Methods in a Program for Predicting the Antioxidant Activity of Winemaking Materials 131
Nadezhda Gallini, Anatoliy Kazak, Marina Rudenko, Yuriy Grishin, Anisa Atik, and Victor Gallini

Composition of Differential Privacy Mechanisms in Iterative Machine Learning: A Comparative Study 136
Saveliy Gordeev, Anton Sinitsyn, and Maria Lapina

Design of Number-Theoretic Transforms Using Diemitko Numbers 149
Saveliy Gordeev, Anton Sinitsyn, and Maria Lapina

Application of Artificial Intelligence in Telecommunications 162
Svetlana Govorova, Egor Govorov, Malvina Karabasheva, and Neelakandan S

Neural Network Method for Solving the System of Differential Equations of the Dynamic Input-Output Balance 171
Natalia Gudieva

Feature Space Optimization and Ensemble Learning Models for Robust Network Intrusion Detection 182
Mohammad Hashir, Mohammad Zayeem, Maria Lapina, and Mohammad Sajid

Comparative Analysis of URL and Hyperlink Features for Machine Learning-Based Phishing Website Detection 196
Jenefa J, Mary Anita E A, Vinodha D, and Vladimir Lapin

Comprehensive Assessment of Ethanol and Extract Content in Wine Products 205
Anatoliy Kazak, Ruslan Timofeev, Nikolay Oleinikov, Marina Rudenko, Anisa Atik, and Anastasia Bobarykina

Analysis of Methods for Detecting Anomalies in Information Systems Based on Containerization Technology 215
Egor Kenkov, Andrey Pavlov, and Mikhail Rusanov

AI-Driven Optimization of Quantum Key Distribution Protocols: a Reinforcement Learning Framework for Adaptive Quantum Cryptography 224
Maksud Khamidov

Conditional Diffusion Model of Image Restoration: Theory, Architecture and Experimental Evaluation 239
Daniil Khomyakov, Vitaly Kudinov, and Konstantin Makarov

Optimizing Alt Text Design through Human-LLM Collaboration 251
Yekaterina Kosova and Kirill Redkokosh

Cascade Classification Algorithm for Malware Detection in Portable Executable Files Based on Static Analysis 263
Alexander V. Kozachok, Shamil G. Magomedov, Andrey V. Kozachok, and Sergei S. Matovykh

Developing the Readiness of Future Teachers to Use Artificial Intelligence in Their Professional Activities 274
Tatyana Kulikova, Natalya Poddubnaya, Olga Fedina, and Prakasha G S

Methods of Identifying Fake News in the Information Space 289
Vitalii Lapin, Alina Bagautdinova, Sergey Fortunatov, Alisa Kushschenko, and Mbuyu Sumbwanyambe

Bias Problems in Recommendation Systems Based on Artificial Intelligence: Modern Approaches to Identification and Elimination 299
Vitalii Lapin, Karim Gediev, Marina Gordova, and Walaa H. Elashmawi

Machine Learning for Detecting Text Classification Attacks 310
Vitalii Lapin, Roman Gubin, Andrey Martynov, Dmitry Savchenko, and Mohammad Shahid

Security Metrics Assessment in Development Pipelines Through the Integration of SecDevOps and Process Mining 323
Maria Lapina, Alina Bagautdinova, Anton Limbo, and Denis Ledyan

Development of an Intelligent System for Adaptive Learning of Foreign Languages Based on AI Using NLP and Machine Learning Methods 330
Maria Lapina, Alina Bagautdinova, Fennie Kemanguluko Mvula, and Vladimir Lapin

ExpLift: Nonlinear Block-Wise Image Compression Via Matrix Exponential Lifting and Low-Rank Approximation 337
Elena Linnik, Ivan Linnik, Fedor Maryasov, and Tatiana Shushara

AI-Assisted Tools in Academic Vocabulary Learning: A Model of Integration 350
Tatiana Marchenko, Liana Vartanova, Olga Kashirina, and Suhel Mustajab

Achieving Technological Sovereignty: The Following Essay Will Provide an Analysis of Machine Vision Market Trends in Russia and Globally for the Period 2020–2030 359
Vladimir Mironenko

Brain Age Estimation Using Optimal Transport-Based Feature Pyramid Fusion (OTFPF) Network 375
N. Banupriya, Jasmine Gilda, and Velmurugan S

Driver Drowsiness Detection System 388
Hishma Naaz, Mohammed Abdul Qadeer, and Vitalii Lapin

Land Use Land Cover Classification 397
Hishma Naaz, Mohammed Abdul Qadeer, and Vladimir Lapin

A Hybrid Deep Learning and Machine Learning Model for Enhancing the Accuracy of Real-Time Traffic Dynamics Forecasting 411
Minh Hieu Nguyen, Dmitry Ilin, and Irina Ivanova

Development of a Hardware and Software System for Agroclimatic Conditions Monitoring and Machine Learning Datasets Generating 419
Nikolay Oleinikov, Yuriy Grishin, Anatoliy Kazak, and Anastasia Bobarykina

Hybrid Cryptographic Approach for File Encryption: Integrating Post-Quantum and Classical Methods 429
Allansius E. Pala, Vinodha D, Mary Anita E A, and Maria Lapina

Problems and Prospects of Using Artificial Intelligence in Education 442
Olga Pankratova, Ekaterina Konopko, and Lusine Bagdasaryan

Assessment of Cyber Resilience of Critical Information Infrastructure to Hybrid Impacts 452
Yulia Parshenkova and Elena Maksimova

Network Traffic Classification Using Complex Machine Learning Methods 461
Tatiana Peleshenko, Dmitry Ardeev, Irina Ivanova, and Mohammad Sajid

Ensemble Neural Networks for Agricultural Tasks: a Comparative Analysis of Models 471
Tatiana Peleshenko, Nikita Boikov, Albina Karaeva, Ivan Borodulin, and Mahfooz Alam

Application of Machine Learning Algorithms in Recommendation Systems 484
Tatiana Peleshenko, Mira Ermolovich, Albina Karaeva, Anton Afanasev, and Alexander Gish

Comparative Analysis of Artificial Intelligence Model Extraction Methods 494
Tatiana Peleshenko, Maxim Martemyanov, Zahid Raza, Ivan Borodulin, and Alexander Gish

The State of Quantum Machine Learning: Architectures, Encoding Strategies, and Generalization Challenges 509
Alicia Theresa Pereira, Sharlene Anna Pereira, Melsa Magdalene Fernandes, and Mary Anita E A

Assessment of the Level of Problem Disclosure in Technical Assignments Prepared Using Large Language Models 517
Karolina Pominova and Hao Hu

BELM: An AI-Based Hybrid Model for Detecting and Mitigating Cyberbullying Against Celebrities on Social Media 527
Boppuru Rudra Prathap, Vandana Reddy, and Maxim Medvedev

AgroInnovation-Transforming Agriculture Through IoT-Enabled Precision Farming for Sustainable Productivity 545
Boppuru Rudra Prathap, Vandana Reddy, and Alexander Zubkov

Unpacking the Knowing-Doing Gap Among Pre-service Teachers Regarding Artificial Intelligence in Education 561
Sanskriti Rawat, Robert Linton Tavis Ashton Bell, Prakasha G S, and Sergey Grigoriev

Application of Artificial Intelligence in Ensuring the Safety of Aircraft Flights by Building a Typical Architecture of the Neural Network of an Aviation Enterprise 576
Nina Romancheva, Leonid Firsov, Dmitriy Melnik, and Ivan Rumyantsev

Regulatory Documents Machine-Understandability—A Stage in the Formation of the Methodological Basis for Digital Production 590
Nina Romancheva, Evgeniy Yurkevich, and Lidiya Kryukova

Information System for Evaluating the Profile of an Information Object 604
Marina Rudenko, Milada Krapivina, Dmitrii Kuhta, and Andrei Rudenko

Advanced Epileptic Seizure Detection Using Self-Supervised Convolutional Neural Network (S-CNN) with the Black Widow Optimization (BWO) Algorithm 619
Neelakandan S, Maria Lapina, Roman Moskotlinov, and Vitalii Lapin

Automatic Speaker Recognition Deep Belief Network (DBN) and Modified Social Spider Optimization (MSSO) Under Noisy and Shouted Speech Conditions 632
Neelakandan S, Prakash M, Velmurugan S, Yulia Andrusenko, and Marina Tokmakova

Electroencephalogram-Based Mental Workload Estimation for Multiple Sclerosis Patients Using Contrastive Self-Supervised Learning and Transformer Encoders 646
Neelakandan S, Sergey Yarushev, and Suresh Manic K

Issues and Challenges in Multimodal Sentiment Analysis 661
Priya Singh, Rajesh Lavania, Mahfooz Alam, Mohammad Shahid, and Vitalii Lapin

Automated Diagnosis of Mental Disorders Based on the Analysis of Speech Biomarkers Using Machine Learning 673
Danil Trofimov, Maria Zapriy, Igor Kotilevets, and Daria Smirnova

Conceptual Model of Value Chain Management of a Digital University Based on Artificial Intelligence 685
Alexander Usanov, Oksana Moshchenko, and Vinodha D

Developing an Information Security Risk Assessment Model Using the CVSS 696
Maria Utenkova, Andrey Komkov, Elena Maksimova, and Oksana Mezentseva

Modified Transformer Architectures for Sports Action Recognition from Pose Sequences 708
Swetha V, Amos Bortiew, Aruna S K, and Stuart Barnes

Automated System for Monitoring a Hero's Image on the VKontakte Social Network Using Artificial Intelligence 718
Ilya Zavyalov, Milada Krapivina, Oleg Gabrielyan, Mikhail Gasparyan, and Ivan Kravchenko

Hybrid Dequantized Machine Learning Models for Information Security Incident Detection 728
Sergei Zuev and Victor Trifonov

Review of Cybersecurity-Aware Machine Learning System for Real-Time Fatigue Detection Based on Physiological and Behavioural Data 735
Munis Xamidov, Otabek Esanov, and Manish Singla

Cross-Analysis of Eye Movements and Yawning States for Enhanced Drowsiness Detection Using Deep Learning 748
M. Xamidov, O. Esanov, and M. K. Singla

Author Index 765

About the Editors

Dr. Maria Lapina is an Associate Professor of the Department of Computational Mathematics and Cybernetics at North Caucasus Federal University, Stavropol, Russia. She received Ph.D. in Mathematics from Stavropol State University. She defended her master's thesis in Information Security. Her research interests include Data Science, Machine Learning, Cybersecurity, Trusted Artificial Intelligence, and Edu-Tech. She has more than 100 research papers. She is an Editorial Board Member for Springer Nature, IGI Global. She has participated in international internships at Sapienza University of Rome (Italy), the University of Ljubljana (Slovenia), the Technical University of Dresden (Germany), and the 2nd ASEF Innovation Laboratory of Higher Education (ASEFInnoLab) at Fudan University in Shanghai, China. She is a finalist of the All-Russian management competition "Leaders of Russia."

Prof. Mary Anita E A a passionate educator and accomplished researcher, with over 36 years of experience in teaching and research in the field of Computer Science and Engineering, specializing in Network Security. She holds a Ph.D. in Computer Science and Engineering from Anna University, Chennai, India. Throughout her academic career she has authored more than 100 research publications in reputed national and international journals and conferences. She has successfully completed two significant research projects and has played a pivotal role in securing funding from various government and professional agencies for organizing academic 3 and research focussed events. A dedicated research mentor, she has guided 12 doctoral scholars to successful completion. She is currently serving as the Associate Dean, School of Engineering and Technology, CHRIST (Deemed to be University), Kengeri Campus, India.

Prof. Mikhail Babenko is the Head of the Department of Computational Mathematics and Cybernetics North-Caucasus Federal University, Russia. He received Ph.D. in Mathematical modeling, numerical methods, and software. He took internships in CICESE Research Center in México; Le Quy Don Technical University, Hanoi, Vietnam in 2015; University of Lorraine, Nancy, France in 2017. Research area: Security, Cryptography, Residue Number System, Neurocomputer Technologies, Digital Signal Processing, Big Data, Internet of Things, Fog–Edge–Cloud Computing, Trust, Uncertainty, Scalable and Reliable Systems, Elliptic Curve, Cryptographic Methods of Information Protection, and Cybersecurity.

Prof. (Dr.) Vijay Anant Athavale has an experience of over 35 years in academics and research. He has published several research papers in indexed journals. He regularly contributes in various international conferences such as Conference Chair, Key Note Speaker, Session Chair, and Organizing Committee Member. He is a reviewer for several indexed journals. He has contributed several chapters in books and handbooks published by IGI Global USA, Taylor and Francis Group, UK, Springer, etc.

His research interests include the Internet of Things (IoT), Artificial Intelligence (AI), Machine Learning, Cybersecurity, and Cloud Computing. He is a member of several international professional bodies.

Dr. Umar Muhammad Modibbo is a Senior Lecturer at Modibbo Adama University, Nigeria, with a Ph.D. in Operations Research from Aligarh Muslim University (2021). A recipient of TETFund and InSc Research Excellence awards, his work focuses on Mathematical Programming, Fuzzy Optimization, and SDGs. He is the President of ORIDSAN and an executive member of AFROS and IFORS. With over 50 publications and 1000 citations, Dr. Modibbo is a leading researcher, international reviewer, and speaker contributing significantly to artificial intelligence, decision sciences, and analytics. Editor-in-chief of Nigerian Journal of Operations Research (NIJOR).

Prof. Andrei Tchernykh is a Full Professor in Computer Science at CICESE Research Center (Centro de Investigación Científica y de Educación Superior de Ensenada), Ensenada, Baja California, Mexico, and an adjunct professor at ISP RAS—Institute for System Programming of the Russian Academy of Sciences, Russia. He gained industrial experience as Supercomputer Design Team Leader in Advance Technical Products Corp and Supercomputer Design Department of Electro-Mechanical Enterprise. He received Ph.D. degree in Computer Science from the IPMCE in 1986. He is a Founding Member of the Mexican Supercomputer Society (RedMexSu), a Regional President of Global Association for Academic Supervision (GAAS) Latin America, and Member of the Mexican National Researchers System SNI level II. He is an editorial board member of journals and served as a guest editor for special issues.

Intelligent, Lightweight and Secure Authentication Scheme for Smart Home

Irfan Alam[1(✉)], Zubair Ashraf[2], Mohammad Shahid[3], and Maria Lapina[4]

[1] Faculty of Computer Studies, Arab Open University, Kuwait City, Bahrain
irfan.alam@aou.org.bh
[2] Department of Computer Science and Artificial Intelligence, College of Computing and Information Technology, University of Bisha, Bisha, Saudi Arabia
ashrafzubair786@gmail.com
[3] Department of Commerce, Aligarh Muslim University, Aligarh, UP, India
mdshahid.cs@gmail.com
[4] Department of Information Security of Automated Systems, Institute of Digital Development, North-Caucasus Federal University, Stavropol, Russia
mlapina@ncfu.ru

Abstract. Artificial intelligence (AI) is making significant contributions to the growing trend of people interacting with the settings in which they live. Smart homes, designed to provide both convenience and automation, are an integral part of today's life. However, security vulnerabilities in smart home systems can cause significant damage to property and human life. The major security problems of Internet of Things (IoT) devices arise due to device authentication and key management. Traditional cryptographic methods might work, but they are often too slow and too complex for simple IoT devices. This paper presents a lightweight authentication technique that uses Shamir's Secret Sharing, cryptographic hash functions (SHA-3), and XOR-based session key generation. The proposed method not only uses Lagrange interpolation to generate pseudonyms for the rebuilt secrets, but also protects the sharing of authentication secrets. The suggested protocol uses hashing and XOR to quickly verify identity while protecting against man-in-the-middle, replay, and key-compromise attacks. The security assessment shows that the protocol works well and is suited to IoT environments with limited resources.

Keywords: Smart Homes · Internet of Things (IoT) · Authentication · Security

1 Introduction

In smart homes, artificial intelligence (AI) algorithms can automatically manage lighting and heating systems and anticipate and perform actions the user will take. This trend is expected to continue in the foreseeable future. Artificial neural networks, multi-agent systems, statistical methods, data compression, and fuzzy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 1–11, 2027.
https://doi.org/10.1007/978-3-032-27834-0_1

logic are some of the AI techniques used in modern smart homes [1,2]. Smart homes are part of the Internet of Things (IoT), making life easier in ways never before possible. However, because they are connected, they also raise privacy and security concerns [3]. As the Internet of Things (IoT) enters mainstream use, the threats of theft, hacking, and even cyber attacks are experienced by the user [4]. This is especially the case for the IoT, which uses many devices that are susceptible to hacking as their means of ensuring privacy are ineffective [5,6].

To lower security risks, you should use strong passwords, encrypt your communications, and keep your software up to date [7]. We also need to be very careful about keeping people's privacy safe in smart homes. It is very important to set rules about who can see data, check who can access it, and limit the collection of private information [8]. Due to the rapid developments in IoT gadgets, robust safety mechanisms are required that incorporate physical as well as technological safety measures in order to minimise the possibility of attack and other forms of system breaches [9,10].

The Internet of Things (IoT) has grown significantly and opened the door for new security issues, specifically in the authentication and management of keys. The use of traditional cryptography concepts, such as Public Key Infrastructure and symmetric cryptography, requires high computational complexity, making it unsuitable for constrained IoT devices [11]. As IoT ecosystems become more complex, it is increasingly important to provide secure, lightweight, and scalable authentication to protect against threats such as replay attacks, man-in-the-middle (MITM) attacks, and device impersonation [12,13]. Figure 1 illustrates typical threats and their solutions to smart homes.

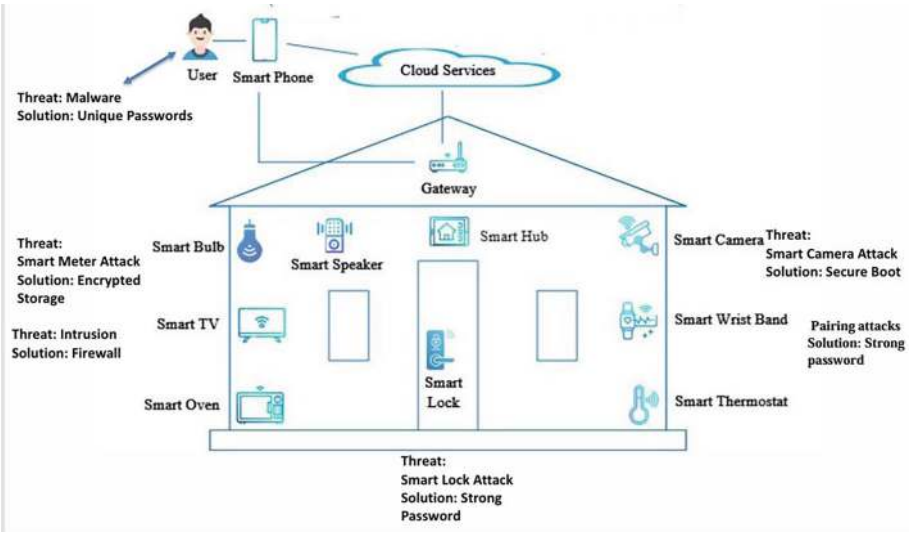


Fig. 1. Smart home threat and its solution

Shamir's Secret Sharing scheme [14] offers an efficient solution for distributed key management by splitting a secret into multiple shares that can be reconstructed with a predefined number of shares. By combining Shamir's Secret Sharing (SSS), cryptographic hash functions, and XOR-based session key generation, this paper introduces a new lightweight authentication protocol for IoT networks. The proposed approach guarantees confidentiality, integrity, and strength against typical cryptographic attacks at a minimal performance cost. The significant contributions of this work are as follows.

- Shamir's Secret-Sharing-based key management: In contrast to the traditional approach of storing static keys, the proposed protocol distributes authentication secrets across several shares, reducing the risk of key loading.
- Interpolation for secure key reconstruction: The use of Lagrange polynomial interpolation allows for efficient key recovery without affecting security.
- Hash along with XOR-based authentication methods: The authentication procedure delivers efficient crypto security via SHA-3 hash algorithms and XOR merge operations.
- resilience to crypto breaches: The protocol uses dynamic session key generation and timestamp-based hashes to prevent MITM, replay, and key manipulation attacks

The rest of the paper is structured as follows. Section 2 provides an overview of the existing work that focuses on authentication schemes for IoT. Section 3 explains the mathematical background that the proposed scheme uses. Section 4 provides the security analysis and performance evaluation. Finally, Sect. 5 consists of concluding remarks.

2 Related Works

Given the extensive use of intelligent gadgets, the Internet of Things (IoT) has become an important area for scientific research. An extensive range of authentication schemes has been designed to provide security for communication processes between IoT networks. Since the schemes are aimed at ensuring that the use of computational resources is minimal, they are particularly effective where resources are limited.

Atzori Antonino et al. [4] investigated the significance of effective authentication methods and the architectural and security issues they present in an important early study. Sicari Sabrina et al. [5] looked into encryption techniques like ECC and RSA. The authors proposed secret sharing and a hash-based authentication method as viable security options for Internet of Things applications. Cloud-based smart home security was addressed by Hossain et al. [9], who suggested encryption schemes that safeguard device communications in such environments. Alam and Kumar [15] suggested a secure protocol for the effective authentication of IoT devices in the cloud. Roman Romero et al. [16] examined the security architecture of cyberphysical systems and proposed a multilayered strategy to mitigate authentication risks. Zhou Wei et al. [11] examined security

and data protection concerns and identified deficiencies in traditional authentication methods, and advocated the development of customized, lightweight encryption for resources-constrained devices. Alam et al. [17] have used secret sharing for a verified multi-secret sharing method and safeguarded secret sharing.

Alfahaid et al. [18] examined AI-based IoT device authentication to improve accuracy and efficiency in IoT environments. Kelaniki et al. [19] examined authentication for IoT devices inside smart cities. Alam [20] introduced IoT authentication techniques adapted for smart city contexts, focusing on scalability, device diversity, and resource limitations. Chen et al. [21] suggested post-quantum cryptography (PQC) for the authentication of IoT devices and secure communication systems. Sun et al. [22] introduced PUF-based authentication for IoT systems, with the aim of improving PUF-based authentication and counter-modeling attacks. Alam et al. [23] proposed attack-resilient authentication schemes are designed with special attention to the low latency, robustness and low storage capacity for Fog-centric IoT device systems.

In this work, we build on these insights and investigate the use of Adi Shamir's secret-sharing mechanism (SSS) in IoT device authentication to reduce exposure to critical vulnerabilities. Our proposed protocol integrates SSS, cryptographic hashing, and XOR-based session constructs to deliver a lightweight yet robust authentication mechanism suited for smart home environments with many devices. Although similar approaches have been studied (e.g., PUF-based or AI-enhanced methods), our focus is on the particular scenario of smart homes and end-to-end delay performance when scaled to tens of devices.

Adversary Model: For any security protocol, there must be a corresponding adversary model to assess its resilience. One of the most widely recognized and fundamental adversary models was introduced by Dolev and Yao in 1983 [24]. This model is widely accepted in the field. It assumes that an adversary can intercept, modify, and decrypt messages if they possess the correct keys. However, they are not capable of conducting statistical or cryptanalytic attacks.

3 Proposed Protocols

A foundational contribution to secure authentication is Shamir's Secret Sharing Scheme (SSS) [14,25], which allows secure key distribution by splitting secrets into multiple shares. The use of SSS in IoT authentication has been explored to reduce key exposure risks and enhance security. The following are the abbreviations of the symbols used in the proposed protocol.

- D_i be the IoT device.
- S be the authentication server.
- K_i be the secret key of D_i.
- $h(x)$ be a cryptographic hash function (SHA-3).
- $\oplus$ denote the XOR operation.
- n be the total number of shares.
- k be the threshold for reconstruction.

– (x_j, s_j) be the shares of K_i.

The proposed protocol has been divided into four steps such as key distribution, authentication request, verification of authentication severity, and secure session key establishment.

Step 1: Secret Sharing-Based Key Distribution

Using Shamir's Secret Sharing scheme (SSS), the key K_i is divided into shares:

$$K_i = f(0) \tag{1}$$

where:

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1} \pmod{p}$$

Each share is:

$$s_j = f(x_j) \tag{2}$$

Step 2: IoT Device Authentication Request

The IoT device reconstructs K_i using Lagrange interpolation:

$$K_i = \sum_{j=1}^{k} s_j \cdot \lambda_j \pmod{p} \tag{3}$$

where:

$$\lambda_j = \prod_{\substack{m=1 \\ m \neq j}}^{k} \frac{x_m}{x_m - x_j} \pmod{p}$$

The device computes:

$$H_i = h(D_i \| T) \oplus K_i \tag{4}$$

and sends (D_i, H_i, T) to the server.

Step 3: Authentication Server Verification

The server retrieves its stored share s_j and reconstructs K_i:

$$H_i' = h(D_i \| T) \oplus K_i \tag{5}$$

If $H_i' = H_i$, authentication is successful.

Step 4: Secure Session Key Establishment

After authentication, both the device and the server compute:

$$SK = h(K_i) \oplus T \tag{6}$$

which expires after $T + 1$.

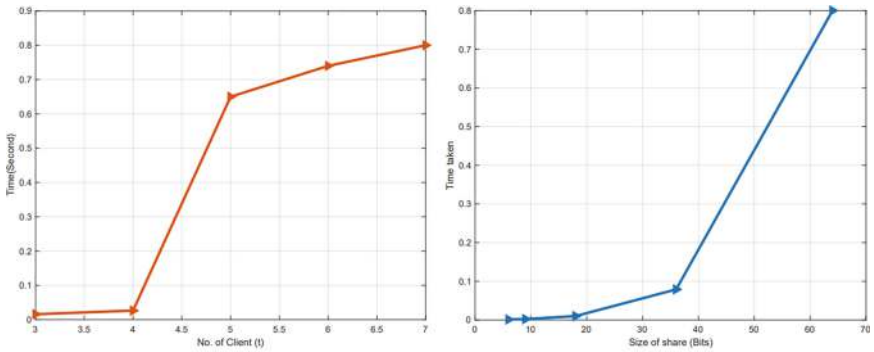
4 Performance Analysis

The proposed authentication protocol is evaluated in this section based on Shamir's Secret Sharing, Hashing, and XOR functions, with respect to computational efficiency, security robustness, communication overhead, and storage requirements. The analysis emphasises the protocol's lightweight nature, security effectiveness, and scalability in IoT environments with limited resources.

4.1 Evaluation of Stealth Share

In Fig. 1, there are two separate analyses based on the share size and the number of participants. In Fig. 2a, the relationship between the number of clients (t) and the time taken in the creation and reconstruction process of the secret is shown. At first, when there are few clients, the time stays slightly above 0 sec, suggesting little effect. Nevertheless, the time increases step-by-step with the number of clients, reaching less than one second with seven clients. Part (b), on the other hand, investigates how the size of the share (in bits) affects the amount of time needed for creation and reconstruction. In this case, the time stays relatively close to zero, indicating that a share size of up to 40 bits has little effect on the procedure.

Each participant must simultaneously manage t + 1 univariate polynomials in order to guarantee authentication free from collusion. As a result, the total storage cost is significantly lower than that of symmetric-key schemes.



(a) Time with respect to no.of client (b) Time with respect to size of share

Fig. 2. Time spent in stealth share in relation to the number of clients and share size

4.2 Computational Efficiency

4.2.1 Key Generation and Distribution The proposed protocol employs Shamir's Secret Sharing (SSS) for secret key distribution. The computational complexity to generate the polynomial in SSS is $\mathcal{O}(k)$, where k is the threshold value for secret reconstruction. Each device requires computing:

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1} \ mod \ p \tag{7}$$

For n number of shares, which takes $\mathcal{O}(k \cdot n)$ time. Compared to RSA-based key exchange ($\mathcal{O}(n^3)$) [11,26], SSS provides a significantly lighter computational load.

4.2.2 Secret Reconstruction Reconstruction of the secret using Lagrange interpolation is as follows:

$$K_i = \sum_{j=1}^{k} s_j \cdot \lambda_j \mod p \tag{8}$$

where:

$$\lambda_j = \prod_{\substack{m=1 \\ m \neq j}}^{k} \frac{x_m}{x_m - x_j} \mod p \tag{9}$$

The interpolation process has a complexity of $\mathcal{O}(k^2)$, making it efficient for IoT devices compared to traditional public-key cryptosystems.

4.2.3 Authentication Process Authentication involves computing:

$$H_i = h(D_i \| T) \oplus K_i \tag{10}$$

The SHA-3 hashing operation requires $\mathcal{O}(1)$ time, and XOR operations are $\mathcal{O}(1)$, ensuring that authentication occurs in constant time, reducing processing delay significantly.

4.3 Communication Overhead

The protocol minimizes communication overhead by:

- **Avoiding Public-Key Cryptography**: Unlike *RSA (1024-bit keys)* and *ECDSA (256-bit keys)*, the proposed scheme operates on *128-bit symmetric key equivalents*, reducing transmission size.
- **Efficient Message Exchange**: The device transmits only (D_i, H_i, T), which is constant size $\mathcal{O}(1)$.
- **Reduced Authentication Requests**: Since secret reconstruction is threshold-based, devices do not need to store full key data, reducing resynchronization overhead.

Three authentication protocols, RSA-1024, ECDSA-256, and suggested SSS+XOR, are briefly compared in Table 1 according to key size, communication overhead, and authentication time. With a time complexity of $(O(n^3))$, RSA-1024 has a high communication overhead and uses the largest key (1024 bits). Despite its moderate overhead and reduction in key size to 256 bits, ECDSA-256 still has a relatively high authentication complexity of $(O(n^2))$. On the other hand, the suggested protocol is more effective and lightweight for resource-constrained environments since it employs a much smaller 128-bit key, achieves low communication overhead, and offers much faster authentication with constant-time complexity $(O(1))$.

Table 1. Comparison of communication overhead

Protocol	Key Size	Comm. Overhead	Auth. Time
RSA-1024	1024 bits	High	$\mathcal{O}(n^3)$
ECDSA-256	256 bits	Moderate	$\mathcal{O}(n^2)$
Proposed (SSS+XOR)	128 bits	Low	$\mathcal{O}(1)$

The proposed protocol requires fewer bits per authentication request, making it bandwidth-efficient for IoT networks. Table 2 shows a comparison of the security features of the suggested SSS-based scheme, RSA, and ECDSA. RSA has only moderate resistance to MITM attacks and does not support lightweight computation. It also provides neither key exposure protection nor replay attack prevention. Although it still lacks key exposure protection and lightweight performance, ECDSA increases security by offering strong MITM resistance and moderate replay attack prevention. On the other hand, the suggested approach offers inherent key exposure protection through secret sharing (SSS), robust MITM resistance, full replay attack prevention, and lightweight computation, which makes it ideal for IoT environments with limited resources.

4.4 Security Analysis

4.4.1 Resistance to Common Attacks :

The protocol resists multiple cryptographic attacks:

- **Replay Attacks**: Authentication requests include a timestamp T, preventing unauthorized retransmissions.
- **Man-in-the-Middle (MITM) Attacks**: The hash-based challenge-response mechanism ensures that intercepted messages cannot be used for impersonation.
- **Key Compromise**: Since authentication secrets are split into shares, no single compromised device reveals the entire secret.
- **Brute Force Resistance**: The SHA-3 hash function increases the computational effort required for key recovery, providing 2^{256} *entropy.*
- **Forward Secrecy**: The session key is dynamically computed as:

$$SK = h(K_i) \oplus T \tag{11}$$

To ensure that past communications remain secure even if future keys are compromised.

4.5 Storage Requirements

4.5.1 Key Storage

- **No centrally controlled Key Storage**: No single entity holds the entire secret as SSS divides shares among devices.

Table 2. Comparative security analysis

Security feature	RSA	ECDSA	Proposed
Replay Attack Prevention	No	Moderate	Yes
MITM Attack Resistance	Moderate	Yes	Yes
Key Exposure Protection	No	No	Yes (SSS)
Lightweight Computation	No	No	Yes

– **Threshold-Based Storage**: Devices can store only k-1 shares, resulting in a storage complexity per device of $\mathcal{O}(k)$.

Table 3 shows that, compared to conventional cryptographic schemes, the suggested SSS utilizing XOR provides a notable increase in efficiency. It is more efficient than both RSA-1024 ($\mathcal{O}(n^3)$) and ECDSA-256 ($\mathcal{O}(n^2)$) because it only requires 128 bits of key storage (distributed among participants) and lowers the secret reconstruction cost to $\mathcal{O}(k^2)$.

Table 3. Memory consumption

Scheme	Key storage	Secret Rec. Cost
RSA-1024	1024 bits	$\mathcal{O}(n^3)$
ECDSA-256	256 bits	$\mathcal{O}(n^2)$
Proposed (SSS+XOR)	128 bits (divided)	$\mathcal{O}(k^2)$

4.6 End-To-End Delays

In this sub-section, the comparative evaluation of the total delay involved in the authentication process, starting from the initiation phase until the termination phase, was performed. The scalability and responsiveness of the authentication schemes are significant and necessary from the perspective of the smart home setting, which, as discussed, consists of more than 10–20 smart devices. On the basis of the analysis, the total end-to-end delay (EED) as calculated for the existing schemes, as referred to in [27–30], equals 0.2384 s, 0.2342 s, 0.2342 s, and 0.232 s, respectively. The total delay for the authentication process for 20 smart devices for the proposed scheme was calculated to be less than 0.01 s, which showcases improved efficacy concerning the response aspect, pointing towards the efficient support of the delay-sensitive smart home applications by the proposed scheme.

5 Conclusion

This paper introduces a novel authentication protocol to improve security, scalability, and efficiency in the Internet of Things (IoT). The proposed protocol uses Shamir's Secret Sharing, along with Hashing and XOR operations, to construct pseudonyms for authentication secrets. Additionally, using Lagrange interpolation guarantees strong key distribution security. The authentication protocol, which is especially designed to improve the security of smart homes, ensures a high degree of security while minimizing the expenses related to processing, storage, and data transmission. A performance investigation revealed that its low overhead and strong resistance to cryptographic assaults make it an appropriate alternative for resource-constrained IoT applications. The concept has the potential to be an excellent match for smart hospitals and companies in smart citie

References

1. Li, K., Li, C., Yuan, X., Li, S., Zou, S., et al.: Zero-trust foundation models: a new paradigm for secure and collaborative artificial intelligence for internet of things. IEEE Internet Things J. (2025)
2. Nachaat, M.: Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. Knowl. Inf. Syst. 1–87 (2025)
3. Zhou, J., Lin, X., Zhang, C.: Security and privacy in the internet of things: challenges and solutions. J. Comput. Sci. Technol. **35**(1), 15–30 (2020). https://doi.org/10.4236/jcc.2024.128016
4. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2017)
5. Sicari, S., Rizzardi, A., Grieco, L.A.: Security, privacy and trust in internet of things: the road ahead. Comput. Netw. **76**, 1–13 (2015)
6. Anwar, T., Ansari, Z.A., Ahmed, R., Ashraf, Z., Khan, G.A.: Comprehensive exploration of sustainable cloud computing, its underlying challenges, and synergy with explainable AI. In: Recent Advances in Computational Intelligence and Cyber Security, pp. 373-381. CRC Press (2024)
7. Irfan, I., Kumar, M.: A novel authentication protocol to ensure confidentiality among the Internet of Medical Things in covid-19 and future pandemic scenario. Internet Things **22**, 100797 (2023)
8. Afzal, S., Bokhari, M.U., Alam, M., Husain, M.S., Khan, M.Z., Ashraf, Z.: LWLCM: a novel lightweight stream cipher using logistic chaos function and multiplexer for IoT communications. PLoS ONE **20**(9), e0330976 (2025)
9. Hossain, M.S., Muhammad, G., Kim, S.: Cloud-based smart homes and iot security: a survey. Computers **7**(3), 33 (2018)
10. Anwar, T., Khan, G.A., Ashraf, Z., Ansari, Z.A., Ahmed, R., Azrour, M.: The combination of blockchain and the internet of things (IoT): applications, opportunities, and challenges for industry. In: Blockchain and Machine Learning for IoT Security, pp. 56–76 (2024)
11. Zhao, K., Ge, L.: A survey on the security of IoT frameworks. Futur. Gener. Comput. Syst. **92**, 15–25 (2019)

12. Rhmann, W., Khan, J., Khan, G.A., Ashraf, Z., Pandey, B., Khan, M.A., Ali, A., et al.: Comparative study of IoT-and AI-based computing disease detection approaches. Data Sci. Manag. **8**(1), 94–106 (2025)
13. Touqeer, H., Zaman, S., Amin, R., Hussain, M., Al-Turjman, F., Bilal, M.: Smart home security: challenges, issues and solutions at different IoT layers. J. Supercomput. **77**(12) (2021)
14. Shamir, A.: How to share a secret. Commun. ACM **22**(11), 612–613 (1979)
15. Alam, I., Kumar, M.: A novel protocol for efficient authentication in cloud-based IoT devices. Multimed. Tools Appl. **81**(10), 13823–13843 (2022)
16. Roman, R., Zhou, J., Lopez, J.: On the security and privacy of cyber-physical systems: challenges and opportunities. Futur. Gener. Comput. Syst. **29**(7), 1376–1386 (2013)
17. Alam, I., et al.: A verifiable multi-secret sharing scheme for hierarchical access structure. Axioms **13**(8), 515 (2024)
18. Alfahaid, A., Alalwany, E., Almars, A.M., Alharbi, F., Atlam, E., Mahgoub, I.: Machine learning-based security solutions for IoT networks: a comprehensive survey. Sensors **25**(11), 3341 (2025). https://doi.org/10.3390/s25113341
19. Kelaniki, S.M., Komninos, N.: A study on IoT device authentication using artificial intelligence. Sensors **25**(18), 5809 (2025). https://doi.org/10.3390/s25185809
20. Alam, I., Kumar, M.: A novel authentication scheme for group based communication for IoT oriented infrastructure in smart cities. 12 December 2022, Preprint (Version 2). Available at Research Square https://doi.org/10.21203/rs.3.rs-1571581/v2
21. Chen, Y.-J., Hsu, C.-L., Lin, T.-W., Lee, J.-S.: Design and evaluation of device authentication and secure communication system with PQC for a IoT environments. Electronics **13**(8), 1575 (2024). https://doi.org/10.3390/electronics13081575
22. Sun, D.-Z., Gao, Y.-N., Tian, Y.: On the security of a PUF-based authentication and key exchange protocol for IoT devices. Sensors **23**(14), 6559 (2023). https://doi.org/10.3390/s23146559
23. Alam, I., Ashraf, Z., Rajpoot, C.S.: Attack resilient authentication scheme for fog-centric IoT device systems. In: Swarm Intelligence, pp. 156-169. CRC Press (2025)
24. Cervesato, I.: The Dolev-Yao intruder is the most powerful attacker. In: 16th Annual Symposium on Logic in Computer Science–LICS, vol. 1. Citeseer (2001)
25. Buchanan, B.: Shamir's secret sharing scheme. Asecuritysite.com. https://asecuritysite.com/shares/go_sss. Accessed 23 Oct 2025
26. Wu, S., Hsu, C., Xia, Z., Zhang, J., Wu, D.: Symmetric-bivariate-polynomial-based lightweight authenticated group key agreement for industrial internet of things. J. Internet Technol. **21**, 1969–1979 (2020)
27. Wazid, M., Das, A.K., Bhat, V., Vasilakos, A.V.: LAM-CIoT: lightweight authentication mechanism in cloud-based IoT environment. J. Netw. Comput. Appl. **150**, 102496 (2020)
28. Challa, S., Wazid, M., Das, A.K., Kumar, N., Reddy, A.G., Yoon, E.J., Yoo, K.Y.: Secure signature-based authenticated key establishment scheme for future IoT applications. IEEE Access **5**, 3028–3043 (2017)
29. Farash, M.S., Turkanovic, M., Kumari, S., Holbl, M.: An efficient user authentication and key agreement scheme for heterogeneous wireless sensor network tailored for the Internet of Things environment. Ad Hoc Netw. **36**, 152–176 (2016)
30. Nyangaresi, V.O.: ECC based authentication scheme for smart homes. In: 2021 International Symposium ELMAR, pp. 5–10. IEEE (2021)

Comparative Analysis of the Effectiveness of Various Machine Learning Models in Content Generation Tasks

Yulia Andrusenko[1(✉)], Ivan Golovin[1], Elena Dombrovskaya[2], and Mary Anita E A[3]

[1] North Caucasus Federal University, Stavropol, Russia
iuandrusenko@ncfu.ru
[2] Financial University Under the Government, Moscow, Russia
[3] Christ University, Bangalore, India

Abstract. This article presents a comparative analysis of modern machine learning models used to generate various types of content–text, visual, and audio. Current architectures are considered, including transformers (GPT-4, LLaMA-3, T5), diffusion models (Stable Diffusion 3, AudioLDM 2), and generative adversarial networks (StyleGAN3). The performance of each model is evaluated using recognized metrics: BLEU, ROUGE, and Perplexity for text generation; Inception Score and Fréchet Inception Distance for images; and Mean Opinion Score and Fréchet Audio Distance for audio. The results demonstrate that the best quality of text synthesis is achieved by transformer architectures, visual synthesis by diffusion models, and audio synthesis by latent diffusion systems. It is concluded that diffusion and transformer approaches form the modern standard for generative artificial intelligence. The results of the study can be used in the construction of multimodal generative systems and in the development of methods for comprehensive assessment of the quality of synthesized content.

Keywords: Machine learning · Generative models · Transformers · Diffusion models · GPT-4 · Stable diffusion · Audio LDM · Text generation · Image generation · Audio generation · Quality metrics · Benchmarking

1 Introduction

In recent years, advances in machine learning methods have radically transformed the approach to digital content creation. Generative AI technologies make it possible to create images, videos, music, and texts from scratch, with quality and expressiveness increasingly approaching human-level performance. These advances are opening new horizons in the creative industries, media, and education [1], while simultaneously raising several questions for those using these methods. The development of pretrained language models has become "a key driver in the transition from specialized solutions to universal generation systems capable of adapting to any text domain." As noted in the paper "Pretrained Language Models for Text Generation: A Survey" (2022) by Li, J., Tang, and co-authors [2], the key issue is selecting the most effective model for a specific content generation task, especially given the abundance of different generative models.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 12–23, 2027.
https://doi.org/10.1007/978-3-032-27834-0_2

The relevance of this research is driven by the rapid growth in the number of generative AI models and the diversity of their application areas. Modern models demonstrate impressive results, but their effectiveness depends on the type of content being created and the specifics of their use. One problem is the lack of a universal approach to assessing the quality of generation, especially across different types of data. Therefore, it is critical to understand which models are most effective at specific tasks and which metrics allow for their effective evaluation. The results of the study can be used in the construction of multimodal generative systems [3]. In financial analytics, generative approaches may complement optimization strategies like those in [4]. Security applications, such as malware detection, also benefit from ML-based generation and feature synthesis [5]. Similar, intrusion detection systems can leverage generative models for data augmentation [6]. Latent diffusion has also been successfully adapted for audio synthesis, as in AudioLDM [7].

2 An Overview of Modern Machine Learning Models for Content Generation

Contemporary machine learning models have shown remarkable potential in content creation, unlocking new opportunities for both creative industries and scientific research. Advances in computational algorithms now allow the generation of text, images, and music entirely without human intervention.

Among these capabilities, text generation has emerged as one of the most explored and widely applied areas of generative machine learning. Historically, recurrent neural networks (RNNs) were the go-to solution for generating text [8]. These networks can process sequences of data, taking into account dependencies between previous and subsequent elements. In practice, this means an RNN can produce sentences where each word is influenced by the words that came before it. Yet, a persistent limitation of RNNs was their difficulty in maintaining long-range dependencies across extended text. To address this, models such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) were developed. They can store and retrieve information over long spans, making them invaluable for dialogue systems, machine translation, and article writing [9, 10].

The real breakthrough came with transformer models. Their ability to efficiently capture dependencies across an entire sequence has established them as the fundamental architecture for language generation, as Vaswani highlighted in “Attention is All You Need” [11, 12]. Among transformers, the Generative Pre-trained Transformer family stands out by considering the full context [13] of the text being generated, rather than focusing only on preceding elements. This capability enables the production of text with strong logical coherence and stylistic consistency, solidifying transformers as the contemporary benchmark for high-quality text generation [14] (Table 1).

Image generation has seen widespread development thanks to generative adversarial networks (GANs) and variational autoencoders (VAEs). A GAN consists of two components: a generator and a discriminator. The generator creates an image based on random noise, while the discriminator attempts to distinguish real images from synthetic ones.

Table 1. Examples of text generators.

Model type	Model examples
RNN	Classic RNN, Elman RNN, Jordan RNN
Long Short-Term Memory (LSTM)	LSTM от Hochreiter & Schmidhuber, Deep LSTM, BiLSTM
Gated Recurrent Unit (GRU)	GRU, BiGRU, Seq2Seq GRU
GPT	GPT-1, GPT-2, GPT-3, GPT-4, ChatGPT

Training these components leads to the generation of realistic images, sometimes indistinguishable from real photographs. GANs are used to create artwork, generate faces, and enhance image resolution. VAEs offer an alternative approach based on probabilistic interpretation of data. This allows not only to generate new images but also to control the style and features of latent vector manipulation, making this model suitable for object synthesis, animation, and image editing. Diffusion models, which gradually remove noise from an image to produce realistic results, have recently gained widespread popularity [15]. These models produce images with an incredible level of detail, numerous individual features, and artistic expressiveness, replicating any painting style on which the model was trained [16, 17] (Table 2).

Table 2. Examples of image generator models.

Model type	Model examples
GAN	DCGAN, StyleGAN, CycleGAN, BigGAN
VAE	Vanilla VAE, β-VAE, VQ-VAE
Diffusion	DDPM, DDIM, Stable Diffusion, Imagen, DALLE 2

One of the most challenging tasks for generative models is generating audio and music content. Among the first effective approaches in this area were WaveNet and SampleRNN, which use deep convolutional networks to predict sound waves with high accuracy. Both models are capable of generating realistic voices and musical fragments, reproducing timbre and intonation characteristics. RNN- and LSTM-based architectures are actively used for music generation, allowing for the modeling of note sequences and harmonies. Each new note is born from the previous one, as if continuing the musical thought and preserving the integrity of the melody. And with the advent of audio models like Jukebox, it became a reality to create compositions with a clearly defined style and recognizable vocal features from rhythmic patterns to harmony and melodic lines. This significantly pushed the boundaries of the use of audio generators and opened up new creative possibilities for them (Table 3).

Table 3. Examples of audio generator models.

Model type	Model examples
RNN/LSTM	SampleRNN, DeepVoice RNN, MusicRNN
Convolutional models	WaveNet, Tacotron, MelGAN
Transformers for audio	Jukebox (OpenAI), AudioLM (Google), MusicLM (Google)

3 Comparison Metrics for Evaluating Model Performance

Modern generative models are already confidently creating texts, images, and music in such a way that it is sometimes difficult to distinguish them from human work. However, at the same time, there is a difficult scientific task of how to objectively assess the quality of such content, if there can be infinitely many possible successful options and each of them is convincing in its own way. Unlike classical classification problems, where the result can be unambiguously compared to a reference, generation requires a multifaceted assessment combining statistical, semantic, and perceptual metrics. Text generation, including summarization, machine translation, and creative writing, is assessed using metrics that capture the degree of match with the reference and the internal consistency of the model. The most common of these are BLEU, ROUGE, and Perplexity, which together encompass three dimensions of text quality: precision, recall, and probabilistic predictability. The BLEU metric measures how accurately the generated text matches the reference text at the n-gram level. The basic idea is that good results are achieved when the model uses similar word sequences as a human. Formally, BLEU is calculated as:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \times \log(p_n)\right)$$

where p_n is the n-gram matching accuracy, w_n is the weight, and BP (brevity penalty) is a penalty for excessive text reduction. BLEU works well for tasks with unambiguous answers–for example, machine translation or paraphrasing [18] – but is limited when evaluating creative and ambiguous texts, where multiple correct variants are possible. In contrast to BLEU, the ROUGE metric focuses on content completeness and is used for summarization or summary generation. The most common variant is ROUGE-L, which uses the length of the longest common subsequence (LCS) between the generated and reference text:

$$ROUGE^{-}L = \frac{LCS(X, Y)}{\max(|X|, |Y|)}$$

where X and Y are the reference and generated texts, respectively. ROUGE-L allows for the preservation of key meanings even if the syntactic structure is altered, making the metric more resilient to paraphrasing. The Perplexity metric characterizes how "natural"

the proposed text is from the perspective of the language model. It is determined by the cross-entropy of the probabilities of predicted tokens:

$$Perplexity = \exp(-\frac{1}{N} \times \sum_{i=1}^{N} \log P(w_i|w_{<i})$$

where $P(w_i \mid w_{((i))})$ is the probability of the word w_i predicted by the model. The lower the Perplexity, the more predictable and grammatical the text is considered. This metric does not require a reference response and reflects the intrinsic quality of the model itself, making it particularly useful for evaluating probabilistic generators such as transformers (GPT, LLaMA). Thus, for text generation, the BLEU–ROUGE–Perplexity triad provides a balance between extrinsic (relative to the reference) and intrinsic (relative to the model) quality assessment, which is critical when comparing architectures with different learning principles and data types. The combination of BLEU, ROUGE, and Perplexity "remains the basic triad of metrics for quantitative and probabilistic quality assessment of language generation," despite the development of more complex models, as noted in Dong, C. et al., A Survey of Natural Language Generation. [19]

Visual models require metrics capable of measuring the realism, diversity, and semantic correspondence of an image to its original description. The most widely used metrics in the scientific literature are Inception Score (IS) and Fréchet Inception Distance (FID), as well as additional metrics such as Precision and Recall, adapted from classical machine learning methods. This metric measures how confidently a classifier (usually Inception v3) recognizes the generated images and how diverse they are. Formally:

$$IS = \exp\left(E_x\left[D_{KL}p(y|x)|p(y)\right]\right)$$

where D_{kl} is the Kullback–Leibler divergence, p(y|x) is the class distribution for the image, and p(y) is the average class distribution over the entire sample. A high IS indicates that images possess recognizable objects (low entropy p(y|x)) and diversity (high entropy p(y)). However, IS does not compare generated images with real ones, which limits the applicability of the metric for tasks requiring photorealism. FID evaluates the distance between the feature distributions of real and synthetic images in the Inception v3 latent space. FID has a high correlation with human visual perception and has therefore become the de facto standard for comparing image generators (GANs, VAEs, Diffusion). For a more nuanced assessment of image quality, the researchers propose considering Precision (realism) and Recall (diversity) as complementary metrics. Precision reflects the proportion of generated images that fall within the real-world data distribution, while Recall measures how well the model covers the entire space of possible real-world images. Together, IS, FID, and (Precision, Recall) allow us to measure the tradeoff between realism and diversity, a key criterion in generative tasks.

Audio signal evaluation is one of the most complex tasks, as sound combines physical characteristics (amplitude, frequency, noise) and the listener's subjective perception. Both subjective metrics (Mean Opinion Score) and objective ones (Fréchet Audio Distance, PESQ, SNR) are used in audio generation. MOS is the average score of how the sound is perceived by live listeners on a scale from 1 (bad) to 5 (excellent). Yes, the indicator is subjective, but that is why it is considered a kind of gold standard in

audio research, because the final quality is ultimately decided by the human ear. In order not to drown in the diversity of opinions, they usually take at least 20 respondents and additionally check that the differences between the options are not random, for example, through ANOVA or Kruskal–Wallis tests. FAD can be perceived as an audio version of FID. This metric measures the distance between the embedding distributions of real and synthetic audio recordings obtained using a pre-trained model like VGGish. In fact, it repeats the formal FID scheme and at the same time is well consistent with MOS, so it is especially useful when you need automatic and scalable evaluation of large datasets. For technical quality analysis, PESQ (Perceptual Evaluation of Speech Quality) is often used–a metric developed for speech quality assessment in telecommunications that takes into account the peculiarities of auditory perception–and SNR (Signal-to-Noise Ratio), a ratio of the useful signal to noise that characterizes the purity of the synthesized signal. Although these metrics are less sensitive to the subjective naturalness of sound, they are useful in analyzing distortions and artifacts at the physical level (Table 4).

Table 4. Metrics for research of text/audio/image generation models.

Content type	Metric	What it measures	Key limitations
Text	BLEU	n-gram match (lexical precision)	Not sensitive to paraphrasing
	ROUGE-L	Semantic completeness (LCS)	Lexically oriented
	Perplexity	Internal predictability of the model	Depends on the evaluation model
Image	IS	Diversity and "recognizability"	Does not compare with real data
	FID	Closeness of feature distributions	Needs large N; sensitive to reference
	Precision/Recall	Realism vs. Coverage	Depends on the thresholds
Audio	MOS	Subjective naturalness	Expensive, variable
	FAD	Proximity of embeddings	Depends on the embedding
	PESQ/SNR	Technical distortions	Do not reflect subjectivity

In summary, each of the metrics discussed measures different aspects of generation quality, and their use should be combined to avoid bias. This systematic interpretation illustrates that no single metric is universal. For example, BLEU and ROUGE provide a formal assessment of match but do not reflect creativity, while MOS captures subjective perception but is not amenable to automation. Therefore, the most robust approach is a hybrid evaluation system combining automated metrics (ensuring objectivity and reproducibility) and expert or user assessments (ensuring consistency with human perception). Evaluation metrics for generative models form the basis for scientific comparison of architectures and enable the transition from subjective quality perception to quantitatively verifiable analysis. BLEU, ROUGE, and Perplexity enable formal evaluation of text generation; IS and FID enable objective comparison of visual synthesis; MOS and

FAD balance subjective and automated audio evaluation. It is the combination of these metrics that provides a multi-level understanding of quality, encompassing accuracy, semantic completeness, realism, structural consistency, and human perception.

4 Selection of Model Evaluation Metrics

In this paper, the comparative analysis will utilize metrics that are currently considered the most effective and robust in their respective fields of application. Since we are considering different types of content–text, images, and audio–the set of evaluation criteria will also differ.

When it comes to evaluating text, the main emphasis is placed on BLEU and ROUGE metrics, as they offer a way to measure how closely generated content aligns with reference samples, taking into account both lexical choices and semantic meaning. Alongside these, Perplexity is also considered, reflecting the model's ability to produce text that feels natural and fluent. This combination makes it possible to combine a rigorous formal assessment with a probabilistic understanding of how naturally and convincingly the model shapes the language.

When it comes to images, the focus is usually on IS and FID. The Inception Score shows two aspects at once: how recognizable the generated objects look and how wide the range of options is. FID, on the contrary, acts more strictly and compares the distribution of features in real and synthetic images, giving a numerical measure of their proximity. Together, these metrics help to understand how the results both look plausible and do not slip into monotony, that is, they reflect both the quality and variety of generation.

Audio quality is best judged with a foot in both worlds. MOS reflects the listener's experience, and FAD puts a hard number on the gap between authentic recordings and generated audio. Used together, they pair human intuition with a clean, reproducible metric.

5 Comparative Analysis of Models

Modern language models based on the Transformer architecture have become a key tool in generating meaningful text comparable in quality to human text. This study analyzed three models: GPT-4 (OpenAI, 2023), LLaMA-3 (Meta, 2024), and T5 (Google, 2020). Each was evaluated using the BLEU and ROUGE-L metrics, which allowed us to quantitatively measure their ability to preserve semantic and syntactic similarity to reference texts. Corpora of news and popular science texts, including summarization and reformulation tasks, were used for the analysis. Each model generated 100 test examples, after which the results were compared with the original reference summaries. The metrics were calculated using standard formulas: BLEU using the geometric mean of n-gram matches, and ROUGE-L using the length of the longest common subsequence between the generated and reference texts. GPT-4 has confidently become a leader both in quality and in working with context, producing a complete and logically assembled text. Its architecture, based only on the decoder, helps to deeply build semantic connections, which is confirmed by BLEU $= 0.64$ and ROUGE-L $= 0.78$. At the same time, the model has a weak point. It often accelerates and generates too much, so when processing

summary data, the result is often overloaded and redundant. LLaMA-3 is almost on par, showing BLEU = 0.61 and ROUGE-L = 0.75, but it looks more disciplined in the sentence structure and makes fewer factual errors thanks to RLHF training. An additional advantage is the low computational complexity, which makes the model especially convenient for solving open research problems. The T5, which works in text-to-text logic, scored slightly less impressive values of BLEU = 0.58 and ROUGE-L = 0.71, but it turned out to be the most stable in terms of final quality and kept the balance between accuracy and reproducibility best. She feels especially confident where the rules are set rigidly and the context is tight, when impeccable wording is more important than vivid presentation (Table 5).

Table 5. Summary of comparison of text models.

Model	Architecture	BLEU	ROUGE-L	Generation features
GPT-4	Decoder-only transformer	0.64	0.78	High coherence and depth of context; tendency to hypergeneration
LLaMA-3	Decoder-only Transformer (RLHF)	0.61	0.75	Structural stability, error tolerance, high efficiency
T5	Encoder-decoder transformer	0.58	0.71	Formal precision, interpretability, predictability of output

The results show that GPT-4 still sets the bar in text generation, combining almost flawless word selection with confident retention of context. LLaMA-3 looks like a practical compromise between quality and resources, so it is convenient to use it for both research tasks and application scenarios. Although T5 is inferior to the leaders in key metrics, it still remains an important tool due to its intuitive logic and ability to adapt to different conditions. As a result, it is worth choosing a model not only based on the numbers in the table, it is important to take into account the cost of calculations, how easy it is to manage the generation and how transparent the final result is.

Modern image generation models have made significant progress. They are already confidently approaching photorealistic quality and are able to build complex visual images with rich semantic content. Three models were considered in this study: Stable Diffusion 3 (Stability AI, 2024), DALL E 3 (OpenAI, 2024), and StyleGAN3 (NVIDIA, 2022). A set of text-to-image tasks with a single text input was used for their comparative analysis. Each model generated 100 images for 10 text prompts of varying complexity. The quality of the results was assessed using the Inception Score (IS), which reflects the diversity and realism of images, and the Fréchet Inception Distance (FID), which measures the distance between the feature distributions of real and generated images. To ensure objectivity, all calculations were performed on the same set of COCO-2017 reference images and at the same resolution parameters (512 × 512). The Stable Diffusion 3 model showed the strongest balance between realism and compositional coherence, achieving IS = 9.2 and FID = 8.7, which reflects both high diversity and a close match to real data distributions. His device, based on diffusion in a hidden space, holds the

semantic connections embedded in the text well and does not "scatter" details, preserving visual integrity, especially where the scene is complex and saturated with elements. DALL·E 3 performed at a comparable level (IS = 8,9, FID = 9,1) and showed itself most strongly in understanding text prompts, especially those where composition is important, for example, "a man reading a book in the rain." A hybrid scheme combining autogenerative and diffusion approaches makes the interpretation of the context more accurate, although sometimes it slightly affects the sharpness of the final picture. StyleGAN3, as a representative of the classic GAN family, scored less (IS = 7.4, FID = 15.3) compared to diffusion models, but it benefits in texture and style control, so it is especially useful where fine-tuning of pose, lighting, or artistic image style is required (Table 6).

Table 6. Summary of comparison of image generation models.

Model	Architecture	IS	FID	Generation features
Stable Diffusion 3	Latent diffusion model	9.2	8.7	Photorealism, compositional accuracy, generation stability
DALL·E 3	Hybrid diffusion-autoregressive	8.9	9.1	Semantic interpretation of text, high context consistency
StyleGAN3	Generative adversarial network	7.4	15.3	Style and pose control, realistic textures, but less variety

The analysis concludes that Stable Diffusion 3 is the most balanced model in terms of visual synthesis quality and image structure stability, providing an optimal combination of realism and diversity. DALL·E 3 demonstrates superiority in interpreting text descriptions and is suitable for creative tasks requiring complex semantic composition. StyleGAN3, despite its comparatively higher FID values, remains effective for applications requiring detailed control over image parameters. Thus, the choice of model is determined by a balance between artistic expressiveness, computational efficiency, and the degree of control over the generation process.

Generative models in audio have evolved from simple speech synthesizers to complex systems capable of creating soundscapes, music, and voice recordings virtually indistinguishable from human ones. This study compares three modern models: AudioLDM 2 (2024), MusicLM (Google, 2023), and VALL-E (Microsoft, 2023). For objective evaluation, two metrics were used: Mean Opinion Score (MOS), a subjective assessment of sound quality perception, and Fréchet Audio Distance (FAD), an objective measure of the similarity of feature distributions between real and generated sounds. Each model was tasked with generating fifty 10-s audio clips based on the same text prompts, covering both speech and musical scenarios. A panel of twenty listeners rated the samples using the MOS scale from 1 to 5, while the FAD metric was computed from embeddings derived via the pre-trained VGGish audio classifier. The evaluation results showed that AudioLDM 2 falls into the golden mean between realism and expressiveness more accurately than others. The metrics MOS 4.6 and FAD 1.21 indicate both pleasant natural perception and minimal discrepancy with the real sound. The hidden spatial diffusion in its architecture helps to make transitions smooth and keep artifacts under control,

and the text module neatly "nails" the sound to a given meaning, preventing it from moving away. MusicLM, designed specifically for music, received MOS 4.4 and FAD 1.37. Yes, it is slightly inferior in acoustic accuracy, but it perfectly builds long musical fragments, maintaining the logic of development. The transformer base confidently captures melodic and rhythmic patterns, which is why the model is especially strong in tasks where coherence over a long distance and a complete composition are important. VALL-E, which uses an autoregressive codec, was released on MOS 4.2 and FAD 1.55 and impressed with how carefully it preserves the individuality of the voice. At the same time, intonation and general naturalness sometimes sound less smooth. But its main trump card is the ability to reproduce timbre and emotional tones with a minimum amount of training data, which makes VALL-E very promising for personalized speech synthesis (Table 7).

Table 7. Summary of comparison of audio generation models.

Model	Architecture	MOS	FAD	Generation features
AudioLDM 2	Latent diffusion model	4.6	1.21	Highly natural sound, precise correspondence to the text description
MusicLM	Transformer-based audio model	4.4	1.37	Excellent musical structure, coherence of long compositions
VALL-E	Autoregressive codec model	4.2	1.55	Accurate rendering of timbre and emotion, but less smooth intonation

A comparison of the models showed that AudioLDM 2 gives the strongest result in terms of overall synthesis quality, successfully combining realism and stability in both speech and music fragments. MusicLM is especially good where you need to build complex melodic structures and keep musical thought at a long distance, so it is best suited for tasks with content that require long-term consistency. The VALL E stands out for its precise imitation of individual voice characteristics, making it a key tool for personalized audio scenarios. As a result, the choice of model is determined by what is more important in a particular task: expressiveness, structural coherence or personalization of sound, but it is the diffusion approaches implemented in AudioLDM 2 that set a new bar in sound generation today.

6 Conclusion

The study conducted a comparative analysis of modern generative models working with various data types–text, images, and audio. Three models representing current trends in machine learning were selected for each category: autogenerative transformers, diffusion architectures, and codec models. The comparison was conducted using rigorous, scientifically recognized metrics: BLEU and ROUGE for evaluating text generation, Inception

Score and Fréchet Inception Distance for images, and Mean Opinion Score and Fréchet Audio Distance for audio. The analysis revealed that the GPT-4 model demonstrated the best results in text generation, providing maximum contextual coherence and lexical accuracy. In the field of visual synthesis, Stable Diffusion 3, distinguished by its stable generation and image realism, leads the way. In the audio segment, the highest scores were achieved by the AudioLDM 2 model, demonstrating the optimal combination of natural sound and structural consistency. Thus, the study confirmed that diffusion and transformer architectures currently form the basis for high-quality content generation across all data types. The choice of a specific model should be determined by the specifics of the task: text models focus on coherence and semantic integrity, visual models on compositional accuracy, and audio models on subjective perception and naturalness. The results obtained can be used in the construction of integrated multimodal generation systems that combine various content types, opening up prospects for further research into unified generative architectures and the evaluation of their cross-modal effectiveness.

References

1. Osorio, P., Jimenez-Perez, G., Montalt-Tordera, J., Hooge, J., Duran-Ballester, G., Singh, S., Radbruch, M., Bach, U., Schroeder, S., Siudak, K., et al.: Latent diffusion models with image-derived annotations for enhanced AI-assisted cancer diagnosis in histopathology. Diagnostics **14**, 1442 (2024). URL: https://doi.org/10.3390/diagnostics14131442
2. Li, J., Tang, T., Zhao, W.X., Nie, J.-Y., Wen, J.: Pretrained language models for text generation: a survey. arXiv preprint (2022). URL: https://arxiv.org/abs/2201.05273
3. Vinodha, D., Mary Anita, E.A., Thirunavukkarasu, P., Geetha, P.: Blockchain-based digital twin for predictive maintenance of machines using machine learning. Blockchain-Based Digital Twins (2025), https://doi.org/10.1201/9781003569046-10/blockchain-based-digital-twin-predictive-maintenance-machines-using-machine-learning-vinodha-mary-anita-padmavathy-thirunavukkarasu-geetha
4. India Mohammad Shahid, Aligarh: Shahid, M., Ashraf, Z., Shamim, M., Ansari, M.S.: A novel portfolio selection strategy using gradient-based optimizer. In: Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Vol. 2 (pp. 287–297). Singapore: Springer Singapore (2021). URL: https://doi.org/10.1007/978-981-16-5348-3_23
5. Maria Lapina, M., Yurdanov, E., Raza, Z., Ivanova, I.: Detection of malware attacks based on machine learning technologies. In: Lapina, M., Raza, Z., Babenko, M., Sajid, M., Zolotarev, V. (eds.), AISMA-2025: International Workshop on Advanced Information Security Management and Applications. AISMA 2025. Lecture Notes in Networks and Systems, vol. 1456. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-07275-7_27
6. Sowmya T., Mary Anita, E.A., Lapina, M.: UK-IDS-machine learning based intrusion detection system for unknown attack detection. In: Lapina, M., Raza, Z., Babenko, M., Sajid, M., Zolotarev, V. (eds.), AISMA-2025: International Workshop on Advanced Information Security Management and Applications. AISMA 2025. Lecture Notes in Networks and Systems, vol. 1456. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-07275-7_38
7. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: AudioLDM: Text-to-Audio with Latent Diffusion". ICML 2023. arXiv preprint (2023). URL: https://arxiv.org/abs/2301.12503
8. Zarrieß, S., Voigt, H., Schüz, S.: Decoding methods in neural language generation: a survey. Information **12**(9), 355 (2021). https://doi.org/10.3390/info12090355

9. Gatt, A., Krahmer, E.: A survey of the state of the art in natural language generation: core tasks, applications and evaluation. arXiv preprint. (2017). URL: https://arxiv.org/abs/1703.09902
10. Sun, B., Li, K.: Neural dialogue generation methods in open domain: a survey. Nat. Lang. Process. Res. **1**(3–4) (2021). URL: https://doi.org/10.2991/nlpr.d.210223.001
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv:1706.03762v7 [cs.CL] (2023). URL: https://arxiv.org/abs/1706.03762
12. Xiao, Y., Wu, L., Guo, J., Li, J., Zhang, M., Qin, T., Liu, T.Y.: A survey on non-autoregressive generation for neural machine translation and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 11407–11427 (2023). URL: https://pubmed.ncbi.nlm.nih.gov/37200120/
13. Zhang, H., Song, H., Li, S., Zhou, M., Song, D.: A survey of controllable text generation using transformer-based pre-trained language models. ACM Comput. Surv. (2023). URL: https://doi.org/10.1145/3617680
14. Sharma, M., Gogineni, A., Ramakrishnan, N.: Innovations in neural data-to-text generation: a survey (2022). URL: https://arxiv.org/abs/2207.12571
15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021). arXiv preprint. URL: https://arxiv.org/abs/2112.10752
16. Zhang, C., Zhang, C., Zhang, M., In So Kweon, Kim, J.: Text-to-image diffusion models in generative AI: a survey. Artif. Intell. (cs.AI); Mach. Learn. (cs.LG). arXiv preprint. (2023). URL: https://arxiv.org/abs/2303.07909
17. Wang, D., Sun, Y.: Exploring various state-of-the-art techniques in diffusion models. 2025S, UCLA CS269 Course Projects. Student project review. (2025). URL: https://ucladeepvision.github.io/CS269-Projects-2025Spring/2025/06/03/student-16-18-diffusionModelSurvey.html
18. Zhou, J., Bhat, S.: Paraphrase generation: a survey of the state of the art. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021). URL: https://aclanthology.org/2021.emnlp-main.414/
19. Anita, E.A.M., Jenefa, J., Vinodha, D., Lapina, M.: Advancements in Sybil attack detection: a comprehensive survey of machine learning-based approaches in wireless sensor networks. AISMA-2024: International Workshop on Advanced Information Security Management and Applications. AISMA 2024. Lecture Notes in Networks and Systems, vol. 863. Springer, Cham (2024).https://doi.org/10.1007/978-3-031-72171-7_7

Deepfakes as a Cyberattack Tool: New Vectors of Social Engineering

Yulia Andrusenko[1(✉)], Dmitry Lukyanov[1], Ivan Golovin[1], Elena Maksimova[2], and Maria Ndapewa Ntinda[3]

[1] North Caucasus Federal University, Stavropol, Russia
iuandrusenko@ncfu.ru
[2] Moscow Technical University of Communications and Informatics, Moscow, Russia
[3] University of Namibia, Windhoek, Namibia

Abstract. This article examines the phenomenon of deepfakes as a new tool of cyberattacks and its impact on the evolution of social engineering methods. The technological foundations of generating synthetic audio and video content, as well as its use for identity substitution in targeted attacks, are analyzed. This article provides an overview of existing approaches to detecting and countering deepfakes, including technical, organizational, and cognitive measure. Particular attention is paid to a comparative analysis of the effectiveness of integrated protection strategies that combine machine learning, regulations, and user awareness. The article highlights the need for a comprehensive approach to ensuring cybersecurity in the face of the growing complexity of multimodal attacks generated by generative technologies and demonstrates the importance of taking into account the human factor when developing adaptive defense mechanisms.

Keywords: Information security · Deepfakes · Artificial intelligence · AI · Cyberattacks · Artificial intelligence technologies · Neural networks

1 Introduction

The rapid development of generative artificial intelligence technologies in recent years has led to the emergence of qualitatively new forms of digital threats that go beyond traditional concepts of cyberattacks. One of the most illustrative and at the same time dangerous manifestations of this process are deepfakes – synthetic audio and video materials created using deep neural networks and capable of imitating real people with a high degree of accuracy. While in the early stages of their development, such technologies were perceived primarily as experimental tools or elements of media culture, they are now increasingly being viewed as fully–fledged tools for cybercriminal activity.

The real danger of deepfakes lies in how effectively they can be woven into social engineering attacks–where the target isn't a system, but a person. Unlike traditional phishing emails or fake websites, deepfakes exploit both sight and sound, making the victim feel they're interacting with someone real. By impersonating the voice or likeness of a trusted individual–like a colleague, friend, or executive – an attacker can quickly

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 24–31, 2027.
https://doi.org/10.1007/978-3-032-27834-0_3

build credibility and dramatically increase the chances of success. As a result, traditional security mechanisms based on routine checks or user vigilance prove ineffective.

The relevance of this problem is increasing in the context of the digital transformation of organizations, the widespread use of remote forms of interaction, and the active use of voice and video communication systems in management and financial processes. Deepfakes are becoming part of targeted attacks aimed at compromising business processes, financial fraud, and unauthorized access to critical information. At the same time, the development of multimodal generative models allows attackers to combine various forms of synthetic content, creating complex attack scenarios, the detection of which requires fundamentally new approaches to ensuring cybersecurity.

2 Theoretical Foundations and Analysis of Related Studies

Our modern understanding of the deepfake phenomenon is shaped by the intersection of research in machine learning, computer vision, speech processing, and cybersecurity. The term "deepfake" has become established as a term for synthetic multimedia content created using deep learning techniques and designed to imitate the appearance, voice, or behavioral characteristics of a specific person. The key feature of deepfakes is not so much the fact of forgery itself, but rather the level of its plausibility, at which the difference between real and generated content becomes elusive not only for an untrained user, but also for automated analysis systems.

From a theoretical perspective, deepfakes are the result of the evolution of generative models, primarily generative adversarial networks and more recent diffusion approaches. These models can be trained on large volumes of multimedia data and reproduce complex statistical relationships characteristic of human speech, facial expressions, and visual appearance. A separate area of research is related to voice cloning, where a synthetic speech model is created based on a limited number of audio recordings, capable of reproducing the intonation, timbre, and rhythm of a specific person. The combination of these technologies has created the preconditions for the emergence of multimodal deepfakes, in which the synthesis of audio and video occurs in a coordinated manner, enhancing the effect of authenticity. Experimental studies show that people are unable to reliably distinguish synthetic speech from genuine speech even with targeted training, making audio deepfakes a particularly dangerous tool for social engineering [10].

In parallel with the development of generative technologies, social engineering methods have also evolved, traditionally viewed as a set of psychological influence techniques aimed at obtaining confidential information or inducing the victim to perform certain actions. Classic social engineering research has focused on textual and verbal manipulation, the use of authority roles, and the creation of situations of urgency. However, the introduction of deepfakes has radically expanded the attackers' toolkit, shifting the focus from abstract messages to imitating specific individuals whose trust has already been formed in the victim's mind.

An analysis of modern scientific papers shows that the use of deepfakes in cyberattacks is considered primarily in the context of specific application scenarios, such as executive scams, attacks on biometric authentication systems, or disinformation campaigns. The ease of access to modern deep learning tools and the rapid development of

generative adversarial networks have made it possible to create deepfakes for the spread of disinformation, revenge, financial fraud, and the subversion of government functions [1]. Researchers note that synthetic content significantly increases the persuasiveness of attacks by reducing cognitive barriers for users and increasing the likelihood of making erroneous decisions. Research into audiovisual deepfakes highlights that consistency between audio and video modalities significantly reduces the effectiveness of both automated detection systems and human perception, increasing the persuasiveness of attacks [9]. However, a significant portion of publications focus on the technical aspects of deepfake detection, including the analysis of generation artifacts, statistical anomalies, and inconsistencies between audio and video streams.

At the same time, the scientific literature increasingly emphasizes that an isolated technical approach to the problem of deepfakes is insufficient given their active use in social engineering. The human factor, trust in visual and auditory evidence, and social contexts of interaction play a decisive role in the success of attacks. Deepfakes are actively used in fraud and social manipulation, requiring a systematic analysis of social risks and regulatory approaches [4]. A number of studies indicate that even with the availability of automated detection tools, the final decision often rests with the individual, whose cognitive and psychological characteristics become a key element of vulnerability.

3 Deepfakes as a New Vector for Social Engineering

Deepfakes have changed the game for social engineering. Instead of casting a wide net with generic phishing emails or scripted calls, attackers can now impersonate someone you know–down to their voice or face–with startling realism. That personal touch builds trust fast and hides deception in plain sight. As AI keeps getting better, deepfakes and other synthetic media are flooding our information spaces, eroding public trust and muddying the line between truth and fiction. This isn't just a tech problem anymore–it's a societal one that demands a broader response [2] (Table 1).

Table 1. Stages of comprehensive counteraction to deepfakes.

Attack phase	Deepfake technologies used	Vulnerability	Method of counteraction
Data collection	OSINT, social networks	Publicity of profiles	Information hygiene policies
Generation	GAN, diffusion models	Lack of control	Synthetic content detectors
Delivery	Messengers, videoconferencing	Channel trust	Verification regulations
Impact	Audio/video	Cognitive distortions	Staff training
Реализация	Financial transactions	Urgency of decisions	Multifactorial checks

One of the most telling markers of how dexeption has evolved is the drowing wave of arracks that don't just mimic a role – they impersonate real people. Using stolen or synthesized voices, faces, and personal mannerisms, attackers now pose as specific individuals – often someone the victim knows, trust, or is accustomed to obeying, whether a manager, a colleague, or even a family member.

What makes these forgeries so dangerous isn't just their visual or auditory realism: it's how well they replicate the subtle rhytms of human interaction: the cadence of speech, familiar turns of phrase, even the way someone pauses or laughs. These details create an uncanny sense of authenticity. The victim doesn't just see or hear a message – the recognize a person.

And that recognition is precisely the trap. Because the interaction feels genuine, the instinct to double-check evaporates. Yet that very insctinct – questioning, veryfying, hesitating – was once the frontline defense against manipulation. Now, it's being quietly disarmed by technology that weaponizex familiarity itself.

A special place in this category of attacks is occupied by audio deepfakes, which ara actively used to compromise voice authentication systems and carry out fraudulent transactions. Voice has traditionally been perceived as a unique biometric feature, which has led to its widespread use in identity verification processes. However, modern speech synthesis methods make it possible to reproduce voice characteristics with high accuracy, including emotional coloring and individual speech characteristics. As a result, voice–based attacks are particularly effective in situations where visual contact is absent and decisions must be made under time constraints.

No less significant is the role played by video deepfakes used in targeted attacks targeting the corporate sector. Using fake video messages or participating in video conferences with a spoofed image allows attackers to influence employees by imitating instructions from management. The use of video deepfakes in corporate communications is considered one of the most dangerous scenarios of targeted attacks, as visual confirmation of identity is traditionally perceived as a sufficient basis for trust [8]. In such scenarios, the presence and visual confirmation of identity significantly reduces the likelihood of critical analysis. In remote work and distributed teams, such attacks become especially dangerous, as visual communication is often considered sufficient proof of the interlocutor's authenticity (Table 2).

Table 2. Deepfake types and attack scenarios.

Type	Channel of influence	Attack type	Sample script	Main vulnerability
Audio	Voice	BEC fraud	Voice substitution	Trust in the voice
Video	Video communication	Targeted attack	Fake Zoom call	Visual confirmation
Multimodal	Audio + video	Complex attack	Funds Transfer Team	Cognitive overload

The most sophisticated and potentially dangerous form of social engineering using deepfakes are multimodal attacks, in which synthetic content simultaneously engages multiple perceptual channels. The coordinated use of audio and video deepfakes creates a coherent illusion of genuine interaction, minimizing the likelihood of detecting inconsistencies. In such attacks, the technical and psychological aspects of impact are closely intertwined, and traditional defense methods focused on analyzing individual modalities lose their effectiveness [14]. Threat analysis of biometric systems shows that speech and image synthesis technologies can effectively bypass voice and facial authentication systems in the absence of multi–factor identity verification [12].

Thus, deepfakes form a new vector of social engineering, in which the key factor in the success of an attack is not the technical vulnerability of a system, but a person's trust in a digital image of a person. This requires a rethinking of existing threat models and a shift in focus from isolated technical measures to a comprehensive understanding of the interaction between technology, human perception, and organizational context.

4 Methods for Detecting and Countering Deepfakes

The development of deepfakes has challenged the scientific and practical communities to find effective methods for detecting them and minimizing the threats associated with their use in cyberattacks. Generative AI enhances social engineering attacks by enabling realistic content creation, high personalization, and automation of attack infrastructure. [3]. The difficulty lies in the fact that deepfake technologies can incorporate multiple types of data: visual, auditory, and human behavioral perception channels, creating coherent multimodal scenarios that are difficult to distinguish from real content. These conditions make traditional methods of protecting information systems and human perception ineffective and require a comprehensive review of cybersecurity strategies.

Technical methods for detecting deepfakes rely on machine learning and deep analysis of multimodal data. On the one hand, algorithms are capable of detecting distortions in created content, discrepancies between image and voice, and anomalies that arise during content creation. On the other hand, the accuracy of these methods is limited by the quality, quantity, and diversity of the training sample, as well as the ability of the models to take into account modern deepfake creation algorithms. Today, detecting deepfakes has become a constant cat–and–mouse game: as generation tools grow more sophisticated, detection methods must keep pace–and often start from scratch with each new deepfake model.

Technology alone isn't enough. Effective defense also depends on clear policies and smart procedures. True security starts with people –not just policies, but practiced habits. It means putting in place ironclad verification steps fot anyone handling sensitive matters, spelling out exactly how confidential information should be treated, and making it routine to double-check messages, requests, or instructions thea raise even a faint eyebrow.

Because no software can catch every lie, and no algorithm can replicate human intuition whe something just feels off. In those moments – and they're becoming more common – sharp eyes, clear thinking, and a culture that encourages speaking up matter more than any firewall.

Today, deepfake detection has become a constant game of cat and mouse: as generation tools become increasingly sophisticated, detection methods must keep up—often

starting from scratch with each new deepfake model. Recent research shows that a combination of technical analysis and employee awareness creates conditions that significantly increase a company's resilience to deepfake attacks. As multimodal attacks become more complex, individual approaches – technical or organizationa – lose their effectiveness. A comprehensive defense strategy, in turn, takes into account both the cognitive abilities of potential victims and thr technical capabilities of the attacker [13].

5 Comparative Analysis of the Effectiveness of Existing Security Approaches

Modern approaches to protecting against deepfake demonstrate a wide range of strategies, but their effectiveness varies significantly depending on how carefully both the technological and organizational–cognitive aspects of the threat are taken into account. Traditional technical methods based on deep learning algorithms are aimed at identifying artifacts and discrepancies between visuals and audio. They are effective in detecting frequently repeated and known deepfake patterns, but face detection challenges when analyzing multimodal models and adaptive scenarios where generation algorithms take into account natural deviations and patterns of human behavior.

Organizational and procedural measures to protect against deepfakes, on the contrary, transform the human factor into an active element of defense. Even the slickest deepfake can be blunted–but only if organizations do more than cross their fingers and run annual training slides. Regular audits and ongoing, hands-on employee education go a long way toward inoculating teams against deception. Yet these alone won't seal the cracks. Human judgment, for all its strengths, is wired to trust its senses: if a voice sounds familiar and a face looks real, we tend to believe it–sometimes to our peril.

Technology can serve as a useful set of guardrails, flagging anomalies or inconsistencies our eyes and ears might miss. But tools are not a cure-all. Real resilience comes from weaving together three strands: smart tech that spots synthetic media, unambiguous protocols that limit how much damage a single breach can do, and people who've been trained not just to notice something strange, but to act on it without hesitation.

In practice, this means detection systems raise the alarm, policies shut down easy exploits, and alert staff connect the dots–often before an attack fully unfolds. None of these pieces works well alone; it's their interplay that builds a defense agile enough to keep pace with increasingly cunning forgeries. No algorithm, no matter how advanced, can shoulder the burden of security by itself. People, processes, and technology must move in sync–or the whole system falters. Organizational practices–like strong verification procedures, staff training, and a culture of vigilance–add a vital human layer of defense. When people are alert and skeptical by default, they're far less likely to fall for even the most convincing fakes. True resilience comes only when smart tools and informed people work together: algorithms catch what they can, and humans fill in the gaps with judgment and awareness. Ultimately, information security becomes a comprehensive process, where each element reinforces the other and forms a holistic strategy to counter deepfake attacks (Table 3).

Of particular interest are the comparative aspects of the effectiveness of different approaches. A comparative analysis shows that technological methods for detecting

Table 3. Comparison of detection and counteraction methods.

Approach	Advantages	Restrictions	References
Technical (ML)	Automation, scalability	Vulnerability to new models	[1, 5, 6]
Organizational	Risk mitigation	Human factor	[3, 11]
Integrative	Maximum stability	Difficulty of implementation	[2, 7, 9]

deepfake attacks, although they can detect most threats, cannot account for the influence of social engineering and the psychological characteristics of the user. On the other hand, methods based solely on staff training and the creation of regulations prove vulnerable to multimodal attacks. The combination of approaches overcomes the shortcomings of individual methods by integrating detection, organizational regulation, and user capacity building into a single system that can adapt to new threats.

Comparative analysis shows that a key indicator of effectiveness is the ability to adapt to new forms of threats. Technological solutions must be flexible and updated to reflect the development of generative models, while organizational measures and personnel training must account for changing social engineering scenarios. Systems that do not integrate these components quickly become obsolete and vulnerable to new types of attacks [15]. In this context, integrated strategies that combine technical, organizational, and human aspects appear to be the most promising, providing robust protection against the increasing complexity and plausibility of deepfakes.

Thus, a comparative analysis of the effectiveness of existing approaches clearly demonstrates that none of the methods alone is capable of coping with the challenges of modern cyberspace. Only the integration of technology, procedures, and the human factor allows us to create adequate protective mechanisms capable of countering new vectors of social engineering and minimizing the risks arising from the active spread of deepfakes.

References

1. Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A.: Deepfakes Generation and Detection: State–of–the–art, open challenges, countermeasures, and way forward (2021), arXiv, https://doi.org/10.48550/arXiv.2103.00484
2. Shoaib, M.R., Wang, Z., Ahvanooey, M.T., Zhao, J.: Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models (2023), arXiv, https://doi.org/10.48550/arXiv.2311.17394
3. Schmitt, M., Flechais, I.: Digital deception: generative artificial intelligence in social engineering and phishing. Artif. Intell. Rev. **57**, 324 (2024). https://doi.org/10.1007/s10462-024-10973-2
4. Ignatiev, A.G., Kurbatova, T.A.: Deepfakes in the Digital Space: Main International Approaches to Research and Regulation (2025), cgitc, https://cgitc.ru/upload/iblock/23f/67eesk6aig8owuk7brnrgwospccquaec.pdf?utm_source=chatgpt.com
5. Chen, J., Yang, M., Yuan, K.: A Review of Deepfake detection techniques. Appl. Comput. Eng. (2025), https://direct.ewa.pub/proceedings/ace/article/view/20955

6. Taeb, M., Chi, H.: Comparison of Deepfake detection techniques through deep learning. J. Cybersecur. Privacy (2022), https://doi.org/10.3390/jcp2010007
7. Singh, S., Dhumane, A.: Unmasking digital deceptions: an integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges (2025). MethodsX, https://doi.org/10.1016/j.mex.2025.103632
8. Bendiab, G., Haiouni, H., Moulas, I., Shiaeles, S.: Deepfakes in digital media forensics: generation, AI-based detection and challenges. J. Inf. Secur. Appl. (2025), https://doi.org/10.1016/j.jisa.2024.103935
9. Hashmi, A., Shahzad, S.A., Lin, C.–W., Tsao, Y., Wang, H.–M.: Understanding audiovisual Deepfake Detection: techniques, challenges, human factors, and perceptual insights (2024), arXiv, https://arxiv.org/abs/2411.07650
10. Mai, K.T., Bray, S.D., Davies, T., Griffin, L.D.: Warning: Humans Cannot Reliably Detect Speech Deepfakes (2023), arXiv, https://arxiv.org/abs/2301.07829
11. Direev, I.D.: Social Engineering in the context of information security. Int. Res. J. (2025), https://doi.org/10.60797/IRJ.2025.153.121
12. Firc, A., Malinka, K., Hanáček, P.: Deepfakes as a threat to a speaker and facial recognition: an overview of tools and attack vectors (2023) Heliyon, https://doi.org/10.1016/j.heliyon.2023.e15090
13. Sar, A., Roy, S., Choudhury, T., Abraham, A.: Zero-Shot Visual Deepfake Detection: Can AI Predict and Prevent Fake Content Before It's Created? (2025), arXiv, https://arxiv.org/abs/2509.18461
14. Qureshi, S.M., Saeed, A., Almotiri, S.H., Ahmad, F., Ghamdi, M.A.Al.: Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media (2025) PubMed/NLM, https://pubmed.ncbi.nlm.nih.gov/38855214
15. Devi, B.T., Rajasekaran, R.: A novel approach to identify deepfake text using social media data. Soc. Netw. Anal. Min. **15**, 112 (2025). https://doi.org/10.1007/s13278-025-01536-6

Development of a Classifier of Motor Transport Based on Neural Networks

Ivan Azarov[1], Elmira Azarova[1], Alexey Dmitrienko[1(✉)], Alexey Kozhevnikov[1], and Mohammed Abdul Qadeer[2]

[1] North Caucasus Federal University, Stavropol, Russia
admitrienko770@gmail.com
[2] Aligarh Muslim University, Aligarh, India

Abstract. The article discusses the problem of classifying cars using deep learning methods. The aim of the work is to develop and compare the effectiveness of various architectures of convolutional neural networks, such as DenseNet, InceptionV3, ResNet and MobileNet, for accurate recognition of vehicle types from images. In the course of the study, the hyperparameters of the models were optimized, which increased their accuracy by an average of 15%. The best result was demonstrated by the DenseNet architecture, which achieved 98.5% accuracy for the Truck class. Software based on the Keras library has been developed that implements the functions of model loading, image preprocessing, and classification in real time. The results of the work confirm the prospects of using deep learning for the tasks of automatic vehicle identification in such areas as video surveillance systems, logistics and road safety.

Keywords: Car classification · Convolutional neural networks · Deep learning · Computer vision · Image recognition · Keras

1 Introduction

The task of data classification is the study of various methods, algorithms, and models aimed at efficiently organizing data into categories or classes based on their characteristics. Within the framework of the article, special attention is paid to the selection of features, the definition of classes, the applied learning algorithms and methods for evaluating the quality of models.

One of the urgent tasks in the field of classification is the development of highly efficient car classifiers, which requires the use of modern machine learning methods, in particular neural networks. Neural networks, due to their ability to identify complex patterns in data, are a powerful tool for solving classification problems, such as recognizing makes, models, or types of cars based on various features such as images, technical specifications, or operating parameters.

The purpose of this work is to develop and improve a car classifier using neural networks, which will improve the accuracy and reliability of solutions in such areas as automatic vehicle recognition. In the course of the work, various approaches to building

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 32–43, 2027.
https://doi.org/10.1007/978-3-032-27834-0_4

neural networks for classifying cars will be analyzed, as well as methods for optimizing and evaluating the effectiveness of such models will be studied.

The object of research in this scientific paper is the processes and methods of classifying motor transport using neural networks. The subject of this research paper is an approach to optimizing the training of neural network models and improving the accuracy of their classification. The purpose of this work is to develop a neural network model that will improve the classification efficiency of motor transport.

Thus, the expected result of this study is the development of an effective classifier of motor transport based on neural networks, providing high accuracy of recognition of various types of vehicles. The model being developed should demonstrate resilience to a variety of input data and have the ability to generalize, which will allow it to be used in real-world conditions for automated analysis and processing of video streams, images or other visual data containing motor transport.

2 Analytical Section

2.1 Classification Methods

Car classification methods using neural networks represent an area of active research and application in modern machine learning.

CNNs are the most common and effective method for classifying images, including images of cars. These networks are capable of automatically extracting features from images at various levels of abstraction, which makes them an ideal choice for the task of recognizing car brands and models [7–9].

In addition to classification, there is also the task of detecting objects in images, which allows you to determine the location and class of an object. Neural network architectures such as Faster R-CNN, YOLO (You Only Look Once), SSD (Single Shot MultiBox Detector) and others are often used for this [5].

The GAN class of neural networks is used not only for classification, but also for generating images of cars. GAN models can create realistic images of cars that can be used to expand the training dataset or to augment existing data.

These methods have demonstrated significant success in classifying cars in images, providing high accuracy and generalizability for different types of cars and shooting conditions. However, their application requires careful hyperparameter tuning and data processing to achieve optimal results.

2.2 Relevance of Image Classification Based on Neural Networks

The relevance of classifying cars by images using neural networks is due to a number of factors that affect various aspects of our lives, from improving road safety to optimizing processes in the automotive industry and developing taxi and carsharing services.

In general, classifying cars by images using neural networks is a powerful tool that finds application in many areas, contributing to improving the quality of life, improving safety and optimizing various processes.

Classification of cars using neural networks has undeniable advantages that make it feasible and effective: high accuracy; generalizability; automation and scalability; adaptability to changes.

However, it is worth noting that classifying cars using neural networks also has some limitations and challenges, such as the need for a large amount of training data, the complexity of configuring models, and computational costs. Nevertheless, given the constant development of learning methods and the growth of available computing resources, the use of neural networks for car classification remains a highly valuable tool with great potential.

2.3 Literature Review

The field of classification using neural networks is actively studied and described in the scientific literature. The article "ImageNet Classification with Deep Convolutional Neural Networks" (Krizhevsky et al. 2012) describes the architecture of the AlexNet convolutional neural network, which for the first time showed amazing results in image classification based on the ImageNet dataset. This breakthrough became the starting point for the use of neural networks in the field of computer vision [1].

The paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" (Simonyan and Zisserman 2014) presents the VGGNet architecture with deep convolutional layers, which demonstrated that even deeper models can provide even better performance in image classification tasks [2].

The study "Going Deeper with Convolutions" (Szegedy et al. 2015) presents the Inception architecture (also known as GoogLeNet), which introduces perceptual blocks to improve the efficiency of using parameters in deep convolutional neural networks [3].

The article "Deep Residual Learning for Image Recognition" (He et al. 2015) describes the ResNet architecture, which introduces blocks with residual connections for training deep neural networks, overcoming the problem of gradient attenuation and allowing efficient training of deep neural networks [4].

The study "YOLO: Real-Time Object Detection" (Redmon et al. 2016) presents the YOLO (You Only Look Once) architecture, which combines the tasks of object detection and classification into a single model, providing high speed and accuracy in real time [5].

The work "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" (Tan et al. 2019) presents a methodology for scaling convolutional neural networks, which allows you to automatically adjust the depth, width and resolution of the model to achieve optimal performance [6].

These works represent a significant part of the extensive research field of classification using neural networks and demonstrate the continuous development of methods and architectures for solving classification problems in various fields, including images, texts, sounds and other types of data.

There are many developments for classifying cars based on neural networks presented in the scientific literature and used in the industry.

ImageNet is an extensive image database designed for training and testing computer vision algorithms. It includes millions of images divided into thousands of categories,

including images of cars. ImageNet is one of the most widely used datasets for developing and evaluating image classifiers. Appropriate labels and categories are provided for car images, which facilitates the development of specialized classifiers. Despite the vastness and diversity of ImageNet, the presence of a large number of categories can be a problem when developing classifiers for specific tasks. Some classes in ImageNet may be too general or not meet the technical requirements, which may require additional data preprocessing or the development of specialized algorithms for filtering or combining classes. This can complicate the development process and reduce the accuracy of classifiers if you do not take into account the specifics of the task when selecting and using data from ImageNet [1].

The open source Vehicle classification classifier using keras demonstrates high efficiency in detecting cars on the road. Its architecture and learning algorithms make it possible to accurately determine the presence of a car in the frame, making it a valuable tool for autopilot systems. However, despite its advantages, this classifier has limitations in its application. It is optimized to work in traffic conditions and may not provide the necessary accuracy in other contexts. Therefore, its use is limited to autopilot systems and is not always suitable for other tasks requiring car recognition in various situations [15].

The Car Model Classification car classifier demonstrates high efficiency in classifying cars. However, its capabilities are limited in terms of recognizing specific car models. The area of classification of specific car models has too small an application area, which can be a problem for many tasks. To maintain high accuracy of the classifier, it is necessary to regularly retrain it when new car models appear. This is due to the fact that modern automobile manufacturers produce new models with regularity, and the classifier must adapt to these changes in order to maintain its effectiveness [16].

The Vehicle Classification Computer Vision Project is the most versatile solution for classification tasks. However, despite its versatility, it does not cover all the classes of cars needed to solve some specific tasks. In such cases, it may be necessary to modify the classifier to include additional classes of cars, which may negatively affect accuracy [17].

These classifiers represent different approaches, each of which has its advantages and disadvantages depending on the specific task and application conditions. It is important to choose the appropriate classifier, taking into account the requirements for accuracy, speed, resources, and other factors.

3 Project Section

3.1 Designing and Analyzing the Results

As part of this project, a classifier program has been developed that uses artificial intelligence methods, in particular neural networks, to automatically classify cars according to various criteria. The aim of the development is to increase the accuracy and speed of vehicle identification with minimal operator involvement. The program can be used in monitoring systems, transport hubs, logistics, parking lots, and security systems.

The program is based on a convolutional neural network (CNN) trained on a set of images of cars. The input information is photos of vehicles taken from the surveillance camera or uploaded manually. The output is a set of predicted car type.

The system is capable of processing images in real time and producing results with high accuracy.

During the tests of training models using different architectures of neural networks, the following results were obtained. Hyperparameter optimization experiments were conducted for each model, during which the best parameters were identified to ensure maximum performance on the validation dataset. The process of configuring hyperparameters included various methods, such as selection using grid search, random search, or Bayesian optimization methods. The results of these experiments made it possible to select the most appropriate combinations of hyperparameters for each model, which improved the accuracy and stability of its operation. An important aspect is that for each neural network architecture, hyperparameter optimization was carried out taking into account the specifics of the task and the type of data, which played a key role in improving its performance (Table 1).

Table 1. The most suitable combinations of hyperparameters.

Model	Learning rate	Batch size	Image size	Optimizer	Regularization
DenseNet	0.0001	64	224 × 224	Adam	25%
Inception V3	0.001	64	224 × 224	Stochastic Gradient Descent	50%
ResNet	0.001	64	224 × 224	Adam	50%
MobileNet	0.0001	64	224 × 224	Adam	–

The results of testing various neural network architectures allow us to draw an important conclusion: the correct setting of hyperparameters can significantly improve the accuracy of the model, increasing its efficiency by about 15% (Fig. 1). Based on the conducted research, it can be confidently stated that the DenseNet architecture has demonstrated the highest accuracy in this task. At the same time, the InceptionV3 and ResNet architectures are somewhat behind the leader, but still show high results, which confirms their competitiveness within the framework of the task under consideration. At the same time, the MobileNet architecture, despite showing the worst results compared to other models, still showed decent performance. It should be noted that MobileNet is designed with an eye to use in conditions of limited computing power, which allows it to maintain high accuracy with relatively low device performance requirements. This aspect is important because tasks that require the use of mobile and embedded systems require models to balance accuracy and computing resources, which makes MobileNet an interesting option for use in such scenarios.

Based on the results of the experiments, it is possible to identify differences in classification accuracy between different classes of objects. The Truck and Pickup classes demonstrated the highest recognition accuracy, which indicates the high efficiency of

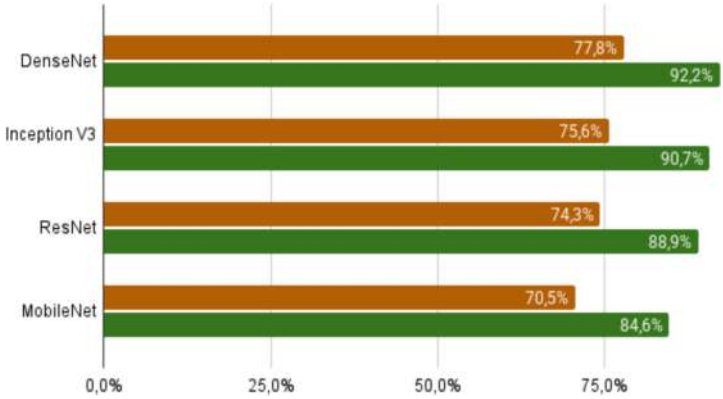


Fig. 1. Graph of model accuracy before and after hyperparameter adjustment.

the model in recognizing these categories of objects (Table 2). At the same time, average results were recorded for the Van and SUV classes, which indicates that there are certain difficulties in classifying objects of these classes, despite the relatively high overall performance of the model. The Convertible, Coupe, Hatchback, and Sedan classes demonstrated the lowest recognition accuracy, which may indicate the presence of specific features in these categories that make it difficult to accurately identify and classify them.

Table 2. Class recognition accuracy.

Class	DenseNet (%)	Inception V3 (%)	ResNet (%)	MobileNet (%)
Pickup	96,4	95,1	92,1	88,8
Convertible	89,4	86,6	85,5	81,8
Coupe	88,8	86,9	84,9	81,2
Hatchback	90,5	89,8	88,7	82,9
Sedan	88,3	85,9	84,8	80,7
Suv	92,6	91,2	89,1	85.0
Truck	98,5	96,9	95,2	90,9
Van	93,1	93,2	91,1	85,5

The maximum recognition accuracy was achieved for the Truck class, which showed 98.5% accuracy when using the DenseNet architecture. This result indicates the high efficiency of the DenseNet architecture in solving the classification problem for this particular class of objects. In contrast, the lowest recognition accuracy was recorded for the Sedan class, which could be due to the lower differentiability of visual features of objects of this class compared to others, as well as the features of the model using the

MobileNet architecture. These data indicate the need for further research and optimization of the model, especially for classes that exhibit lower recognition accuracy, in order to improve its versatility and applicability in real classification tasks.

3.2 Description of the Domain Implementation

Before starting training a convolutional neural network, it is necessary to carefully prepare the appropriate dataset, which will serve as the basis for effective training of the model. This process includes several key steps. First, you need to determine the number of classes that the various categories of images will belong to. Next, it is necessary to classify each image according to its belonging to one of these classes, ensuring the correctness and high quality of annotations, since the accuracy of the network directly depends on this. In addition, it is important to take into account the diversity of data so that the training sample represents all possible variations in real conditions. After classifying the images, the data obtained should be divided into three parts: the training set, the validation set, and the test set. This division is critical in order to ensure both correct training of the model and the possibility of its objective evaluation on data that is not involved in the learning process, which avoids overfitting and increases the overall generalization ability of the network.

```
train_dataset=image_dataset_from_directory(r'C:\Us
ers\Lunex\Desktop\CARS\car_data2',
      subset='training',
      seed=42,
      validation_split=0.1,
      batch_size=batch_size,
      image_size=image_size)
```

The Keras library has a convenient and efficient mechanism for connecting various neural network architectures, which provides ample opportunities for their flexible configuration. This feature allows the user to quickly integrate pre-prepared models or use standard architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), as well as more complex hybrid models. At this stage, weights are also connected, pre-trained based on the ImageNet dataset, which allows you to speed up the learning process and increase the accuracy of the model. In addition, Keras provides extensive interfaces for configuring architecture parameters such as the number of layers, activation functions, initialization parameters of weights, as well as optimization and regularization parameters. As a result, using Keras, it is possible to effectively adapt the neural network architecture to a specific task, minimizing setup time and improving the accuracy of the solution.

```
DenseNet121_net=DenseNet121(weights="imagenet",
    include_top=False,
    input_shape=(224, 224, 3),
    classes=8 )
    DenseNet121_net.trainable = True
DenseNet121_net.summary()
```

It is possible to partially train the weights of the model. Below is the code that allows you to train only that part of the model that is located after a certain layer.

```
trainable = False
for layer in DenseNet121_net.layers:
    if layer.name == 'conv5_block1_0_bn':
        trainable = True
    layer.trainable = trainable
```

To correctly perform the task of classifying objects, an output layer must be added to the previously connected neural network architecture. This layer plays a key role in the classification process, as its function is to transform the information obtained in the previous stages into the final class labels. Usually, the output layer includes neurons, the number of which corresponds to the number of classes, and also uses an activation function corresponding to the type of task (for example, the Softmax activation function is often used for multiclass classification).

```
model = Sequential()
model.add(DenseNet121_net)
model.add(Flatten())
model.add(Dense(256))
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(8,))
model.add(Activation('softmax'))
```

After adding the output layer, the model is compiled and trained using the Keras library. At the stage of model compilation, key components such as the loss function, optimizer, and metrics are established, which will be used to evaluate the effectiveness of the model in the learning process. The loss function uses categorical crossentropy, which is the standard for multiclass classification problems where class labels are represented by integers. This loss function effectively measures the discrepancy between the predicted probabilities and the true class labels.

Also, at this stage, an optimizer is selected and the appropriate learning rate is set, which is necessary for stable convergence of the model. This prevents steps that are too large, which can lead to instability in the learning process.

Accuracy is used as a metric to evaluate the model's performance, which measures the percentage of correctly classified objects relative to the total number of objects. This gives a clear idea of how accurately the model solves the classification problem.

After compilation, the model training stage begins. The model is trained on a training dataset (train set), while a validation dataset (validation set) is used to monitor the learning process and prevent overfitting. The learning process continues for a predetermined number of epochs, which allows for a gradual improvement in the quality of the model.

```
model.compile(loss='sparse_categorical_crossentropy',

optimizer=Adam(learning_rate=0.0001),
              metrics=['accuracy'])
history = model.fit(train_dataset,

validtion_data=validation_dataset,
              epochs=epochs_num,
              verbose=2)
```

After the model training process is completed, its accuracy is evaluated on a test dataset that was not used during the training process. This stage is crucial for testing the generalizing ability of the model, that is, its ability to correctly predict results based on new, previously unknown data. The test dataset is usually allocated at the stage of preliminary data preparation and is used exclusively to evaluate the quality of the model after its training. Various metrics can be used to evaluate accuracy, depending on the type of task: for classification tasks, it can be accuracy, for regression tasks, it can be mean square error or other metrics. This stage allows you to identify possible overfitting of the model, as well as to give an objective assessment of its performance on real data. After evaluating the accuracy, the trained model must be saved for future use in the classifier.

```
scores = model.evaluate(test_dataset, verbose=1)
accuracy = round(scores[1] * 100, 4)
model.save(f"{folder}/model{name}.keras")
```

To implement the task of classifying cars using the model developed in the previous section, it is necessary to create a classifier. The main purpose of the classifier is to process input data using a trained model and then output to the user the name of the class to which a particular car belongs.

Initially, the model is loaded. At the first stage, the classifier loads the model saved in Keras format. This allows you to use the pre-trained weights and network architecture for further image analysis. Next, an image of the car is loaded, which will be used for classification, and the image is adjusted to the required format corresponding to the input data on which the model was trained. This may include resizing, normalization of pixel values, and other preprocessing operations that ensure the correct interaction of the image with the model. After the image is adjusted to the desired appearance, it is passed to the model, which performs data processing and generates a forecast based on its internal parameters.

The processing result is presented as an array of probabilities, where each element of the array corresponds to the probability that the image belongs to one of the possible classes. The classifier selects the element of the array with the highest value, and this element indicates the index of the class to which the image most likely belongs. The index obtained in the previous step is used to determine the name of the corresponding class. This index is compared with an array containing the names of all available classes, and based on this, the name of the class to which the car belongs is output. As a result of the classifier's operation, the user is provided not only with the class name, but also with

its identifier, which makes it convenient for both humans and the system to perceive and interpret information.

Here is a visualization of the classifier using the example image (Fig. 2) of the Van class. An image is provided as input, and the user receives 3 options for determining the class at the output: an array of the probability of belonging to the class, the class number, and its name (Fig. 3).

Fig. 2. The original image.

```
2025-06-03 12:30:18.427318: I tensorflow/stream_executor/cuda/cuda_dnn.cc:384] Loaded cuDNN version 8100
2025-06-03 12:30:19.574101: I tensorflow/stream_executor/cuda/cuda_blas.cc:1614] TensorFloat-32 will be used
1/1 [==============================] - 3s 3s/step
Вероятность принадлежности классам: [[7.9088015e-05 3.1223986e-05 1.4612115e-07 2.3046192e-05 1.2304031e-06
  8.2418395e-05 2.2278311e-05 9.9976057e-01]]
Номер класса: 7
Название класса: Van
```

Fig. 3. Output of the classifier result to the console.

Based on the data obtained, it can be concluded that the classification of the image was successful and the answer received is correct because the highest probability of belonging to the image (9.9976057e-05) is class number 7, which corresponds to the Van class to which the original image belongs.

Thus, the developed classifier is a system that receives an image of a car as an input and provides information on whether the image belongs to one of the classes, expressed as the name of the class and its identifier. This solution is essential for efficient data processing and further analysis, both for users and automated systems.

4 Conclusion

This study was aimed at analyzing existing methods and tools for classifying motor vehicles, as well as studying image classification algorithms using neural networks. As part of the work, classification models were selected, developed and tested, and a specialized classifier for road transport was developed.

The study examined various approaches, technologies, and tools used to classify car images, as well as analyzed the effectiveness of various neural network architectures, activation functions, and optimizers used to evaluate and improve models.

The results showed that the choice of architecture, activation function, optimizer, and hyperparameters has a significant impact on model performance. The developed and tested models showed an improvement over the base model, achieving high accuracy.

The work provided data on image classification methods for motor vehicles, demonstrating the significant potential of using neural networks to solve car classification problems. The project based on the use of neural networks has commercial potential and can be implemented using the Python programming language and the Keras library.

A detailed comparison of different models was carried out in order to choose the optimal one for this task. The software implementation of these models, including the integration of optimized solutions into the software system, is also investigated.

In conclusion, the developed model was tested, its significance for the study was evaluated, and the test results were reviewed. The results of the work highlighted the importance of hyperparameter optimization and demonstrated the process of developing and testing models for image analysis using neural networks. The conclusions and results of this study may be of interest to specialists working in the field of image classification and neural networks.

References

1. ImageNet classification using deep convolutional neural networks [Electronic resource] – URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
2. Very deep convolutional networks for large-scale image recognition [Electronic resource] - URL: https://arxiv.org/abs/1409.1556 (Date of publication 04/09/2025)
3. We go deeper with the help of convolutions [Electronic resource] – URL: https://arxiv.org/abs/1409.4842
4. Deep residual learning for image recognition [Electronic resource] – URL: https://arxiv.org/abs/1512.03385
5. YOLO: Object detection in real time [Electronic resource] – URL: https://pjreddie.com/darknet/yolo/
6. EfficientNet: Rethinking model scaling for convolutional neural networks [Electronic resource] – URL: https://arxiv.org/abs/1905.11946
7. Convolutional neural networks (CNNS): A complete guide for beginners [Electronic resource] – URL: <url> https://towardsdatascience.com/convolutional-neural-networks-cnn-a-comprehensive-guide-for-beginners-fd3c134cd6d0
8. Advantages and disadvantages of convolutional neural networks [Electronic resource] – URL: https://arxiv.org/abs/1708.08296

9. Pros and cons of convolutional neural networks [Electronic resource] – URL: https://medium.com/@jonathan_hui/convolutional-neural-networks-the-pros-and-cons-1e0b0c2c6e0f
10. Advantages and disadvantages of fully connected neural networks [Electronic resource] – URL: https://arxiv.org/abs/1801.07648
11. Ultra-deep convolutional networks VGG (VGGNet) [Electronic resource] – URL: https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/
12. ResNet (34, 50, 101): "residual" CNNs for Russian–speaking users [Electronic resource] - URL: https://neurohive.io/ru/vidy-nejrosetej/resnet-34-50-101/
13. Googlenet [Electronic resource] – URL: https://paperswithcode.com/method/googlenet
14. Effective network [Electronic resource] – URL: https://paperswithcode.com/method/efficientnet
15. Classification of vehicles using keras [Electronic resource] – URL: https://www.kaggle.com/code/fatnabarkouch/vehicle-classification-using-keras/notebook
16. Classification of car models [Electronic resource] – URL: https://www.kaggle.com/code/vedraiyani/car-model-classification

Research and Development of a Software Module for Analyzing the Preferences of Online Store Users

Ivan Azarov[1], Elmira Azarova[1], Marina Tokmakova[1(✉)], Magomed Abdurazakov[2], and Umar Muhammad Modibbo[3]

[1] North Caucasus Federal University, Stavropol, Russia
marinatokmakova175@mail.ru
[2] Russian Academy of Education, Moscow, Russia
[3] Modibbo Adama University, Yola, Nigeria

Abstract. Modern online stores face the problem of processing large amounts of unstructured data, including text user reviews, product descriptions and their characteristics. Traditional analysis methods based on manual processing or simple statistical approaches do not allow us to effectively identify hidden patterns in this data, which reduces the quality of personalization of recommendations and limits the possibilities for improving the service. The article provides research and development of a software module for analyzing the preferences of online store users based on clusterization of text reviews, categorical characteristics of products and numerical ratings. A comparative analysis of clustering algorithms (K-Means, Agglomerative Clustering, DBSCAN) was performed, which revealed the advantage of the DBSCAN algorithm (Silhouette Score = 0.514; Davies-Bouldin Index = 0.934). The developed module is implemented in Python using modern machine learning libraries (scikit-learn, Sentence Transformers, UMAP) and includes the stages of text preprocessing, embedding generation, feature combining, and cluster analysis. The module is integrated as a REST API using FastAPI, which allows it to be used in existing e-commerce systems. The solution is focused on improving the accuracy of recommendation systems, optimizing the product range and improving the customer experience.

Keywords: Clustering of Data · User Preference Analysis · Online Store · Natural Language Processing · Recommendation System · DBSCAN · Machine Learning · E-commerce

1 Introduction

Modern online platforms are forced to compete for the customer's attention among many competitors. Improving the recommendation system allows you to increase customer satisfaction, increase the average receipt and reduce the cost of marketing activities. Thus, the successful integration of the developed module will become an important competitive advantage. Online stores face the problem of processing large amounts of unstructured

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 44–57, 2027.
https://doi.org/10.1007/978-3-032-27834-0_5

data, including text user reviews, product descriptions and their characteristics. Traditional analysis methods based on manual processing or simple statistical approaches do not effectively identify hidden patterns in this data, which reduces the quality of personalization of recommendations and limits the possibilities for improving the service [1].

The modern e-commerce market is characterized by high competition, which requires online stores to introduce personalized approaches to customer interaction [2–4]. One of the key problems in this area is the lack of accuracy of recommendation systems, especially when working with heterogeneous data such as text reviews, product descriptions, and categorical features (color, size, rating).

The relevance of the study is due to the high competition in the e-commerce market and the need to introduce personalized approaches to user interaction. Modern recommendation systems often face the problem of low accuracy, especially when processing heterogeneous data such as text comments, product descriptions, and categorized attributes (color, gender, size, rating). The analysis shows that the use of traditional clustering methods is not effective enough in this context. The research aims to solve this problem by developing a software module capable of effectively analyzing user preferences based on clustering of textual and categorical data.

Object of research: online store users and their behavioral data, including text reviews, product descriptions, product names, as well as categorical features (COLOR, GENDER_TOKEN, MATCHINGSIZE) and ratings (MARK).

Research subject: methods and algorithms for clustering text and categorical data to analyze user preferences and improve the effectiveness of recommendation systems.

The purpose of the study: to develop a software module for clustering data from users of an online store, which allows identifying groups of similar reviews and products, which helps to improve the personalization of recommendations.

2 Literature Review

One of the main problems of the modern market is the low accuracy of personal recommendations due to the lack of methods for analyzing heterogeneous data. The main sources of this complexity are the volume and complexity of text data, the variety of features, and the specificity of the Russian market.

An analysis of research in the field of data analysis and machine learning shows that existing approaches are not able to effectively solve the problem of multicomponent analysis, since most of the proposed solutions focus on one aspect (for example, text-only classification or simple filtering of goods by price). This approach limits the depth of understanding of user needs and does not allow us to identify useful patterns in data arrays.

One of the most common problems is the content of spelling errors in reviews (15–20% of cases), slang and informal expressions [5]. 43% of product descriptions contain repetitions and marketing cliches.

The heterogeneity of the data also causes difficulties, the combination of numerical ratings (mark), categorical features (color, gender_token) and text requires complex

processing methods [6]. Only 28% of platforms effectively integrate all types of data into recommendation systems [1].

The problems of existing solutions do not help, for example, collaborative filtering does not work for new users (cold start problem) [7]. Content-based methods often ignore semantic links between products [8].

Despite the emergence of advanced technologies such as deep learning and graph networks, many projects do not fully apply them, missing out on potential benefits. Thus, most modern projects are limited to the superficial application of standard methods such as k-means or linear regressions, although they do not cope well with large and multi-part datasets [6, 9, 10, 24].

In addition, the libraries and frameworks used, such as PyTorch and TensorFlow, demonstrate high potential for modeling complex relationships within data, but so far they remain poorly integrated into routine analysis processes in the Russian market segment.

Thus, an urgent task is to adapt and modernize existing methods and algorithms for efficient processing of text and numerical data specific to online stores in order to improve the quality of personal recommendations and overall user satisfaction with online store services. Solving this problem will provide companies with significant advantages in the face of growing competition in the e-commerce market.

Despite the existence of a number of publications covering certain aspects of this problem, there is no holistic approach in the available literature that takes into account the totality of the requirements of modern e-commerce.

Next, the main research directions related to the task under study are considered.

The k-means method is convenient due to its simplicity and high computational speed, which makes this method an attractive choice for a large number of applications [11]. However, the method requires a predefined number of clusters and is sensitive to the initial initialization of centroids.

Hierarchical clustering allows you to build a cluster tree, providing more freedom in choosing the level of aggregation [12]. But its performance deteriorates as the size of the dataset increases.

DBSCAN is based on identifying areas of high concentration of objects and is able to find clusters of various shapes [10]. Its disadvantage is the sensitivity to the choice of the parameters ε and minPts.

Despite the wide range of methods, none of them by itself provides sufficient accuracy when processing mixed data (texts + categorical features) necessary for accurate personalization.

Considerable attention is paid to modern approaches to the representation of text in the form of vectors (embeddings).

Word2Vec and Glove create low-dimensional vector spaces where similar words are located close to each other [13]. However, they are less effective for long sequences and depend on the context of the sentence.

BERT and its transformer variations with attention provide deep processing of the meaning of the text, allowing to extract rich semantic information from long fragments of text [14]. Such models allow us to obtain full-fledged vectors of whole sentences, rather than individual words.

Nevertheless, the use of powerful BERT-type models involves high resource requirements and long training time. One of the promising directions is the use of hybrid approaches that integrate text embeddings with numerical and categorical features.

A number of authors point to the high performance of DBSCAN in detecting non-standard clusters and resistance to noise data [10, 15–17]. However, setting the parameters remains a difficult task that requires additional study.

Some scientists propose modifications of DBSCAN aimed at automatically determining the optimal radius (ε) and minimum cluster size (minPts) [18]. However, there are very few practical examples showing the successful application of such modifications in the context of online commerce.

The most important part of the study is not only the identification of clusters, but also their meaningful interpretation and practical application. Work in this field focuses primarily on technical issues of data analysis, leaving open the question of providing interpretable conclusions for businesses [1].

Most studies use a standard approach to assessing the quality of clustering using measures such as the Silhouette Score and the Davies-Bouldin Index [17]. However, specific scenarios for the application of clusterization results in commercial systems have been poorly studied.

Although there are studies confirming the usefulness of data clustering for analyzing consumer preferences [1], real-world cases of successful implementations of such systems in large online stores are rare. Many Russian trading platforms still rely on outdated audience segmentation strategies, which negatively affects the level of personalization of services.

Thus, the review confirms the lack of ready-made solutions that fully solve our task. Our project is aimed at developing a universal and scalable approach that is suitable for the needs of Russian online stores, and allows us to effectively analyze user behavior and improve the quality of recommendations provided.

3 Materials and Methods

3.1 Modeling

The module is implemented in Python using modern machine learning libraries. The paraphrase-multilingual-MiniLM-L12-v2 model was used for text processing, which showed the best results for Russian-language data [19]. Integration with MLflow made it possible to organize model version control and experiment tracking.

The data set uses real data from the Wildberries online store, containing about 150,000 records [20]. The following problems were identified and fixed during the tests:

- uneven distribution of categorical features;
- the presence of outliers in numerical estimates;
- encoding errors in text data.

The main objects of research are: text data, such as user reviews (text), product names (name), their descriptions (description); categorical features, such as color, gender (gender_token), size matching (matchingSize); numerical indicators–product ratings (mark) from 1 up to 5.

Based on the analysis, the research task is divided into the following subtasks:

1. To develop a method for preprocessing text data, taking into account the peculiarities of the Russian language (lemmatization, removal of stop words) using the pymorphy3 and nltk libraries.
2. Implement three clustering algorithms: K-Means for basic data separation; hierarchical clustering for detecting nested structures; DBSCAN for detecting clusters of arbitrary shape and noise points.
3. Evaluate the clustering quality using metrics: Silhouette Score to measure cluster compactness; Calinski-Harabasz index to assess separability.
4. Visualize the results.

3.2 Justification of the Choice of Development Tools

The choice of the technological stack for the implementation of the clustering module of the online store data was carried out taking into account the specifics of the information being processed and the system performance requirements. As a result of the analysis of alternatives, an optimal set of tools was formed to ensure an effective solution to the tasks set.

A special feature of working with the data used is the requirement to preserve semantic relationships in texts while taking into account numerical and categorical characteristics. The main criteria for selecting development tools included:

– natural language processing support for Russian–language texts;
– availability of implementations of clustering algorithms;
– multidimensional data visualization capabilities;
– easy to deploy and scale;
– compliance with licensing requirements;
– community availability and documentation.

Special attention was paid to the compatibility of the tools with each other and the possibility of their integration into a single data processing pipeline.

Three main platforms for data analysis were considered: Python ecosystem, R ecosystem, proprietary solutions (IBM Watson, Google AutoML).

Table 1 shows a comparison of the performance of the platforms under consideration for clustering tasks. Python solutions have shown the best results in terms of speed/quality of data processing.

Bengio's research [21] confirms that Python has become the de facto standard for machine learning tasks due to the balance between performance and ease of development.

The Pandas library is selected for working with tabular data. Alternative solutions (for example, Vaex or Modin) were rejected due to the instability of working with categorical features and limited support for preprocessing algorithms.

The core of the system is based on the scikit-learn library, which provides proven implementations of clustering algorithms, tools for evaluating the quality of models, and data preprocessing mechanisms (StandardScaler, OneHotEncoder).

The choice of DBSCAN as the main algorithm is due to its advantages over K-Means and hierarchical clustering. The scikit-learn library contains an optimized implementation of this algorithm with support for sparse matrices.

Table 1. Platform comparison by key criteria.

Criteria	Python	R	Proprietary solutions
Russian-speaking NLP	5	2	3
Clustering algorithms	5	4	4
Data visualization	4	5	3
Ease of deployment	4	2	5
License purity	5	5	2
Community and documentation	5	4	3

NLP text processing tools used: Sentence Transformers (paraphrase-multilingual-MiniLM-L12-v2 model) to generate embeddings; razdel library for tokenization of Russian texts; Pymorphy2 for lemmatization. The Sentence Transformers model was chosen after a comparative analysis of alternatives (Word2Vec, FastText, BERT) conducted by Reimers [22]. Its key advantage is the ability to preserve semantic connections in multilingual texts with relatively low computational costs.

To analyze the results, the following methods are used: Matplotlib (basic graphs and diagrams); Seaborn (statistical visualization); UMAP (dimensionalization for visualization of clusters). UMAP has shown better results than t-SNE in multidimensional data visualization tasks, which is consistent with McInnes' research [23]. This is especially important for interpreting clustering results.

For industrial operation of the module, the following are used: FastAPI (REST API creation); Docker (solution containerization); MLflow (model lifecycle management). FastAPI was chosen due to its high performance (based on Starlette) and automatic documentation generation. Testing has shown that it handles up to 1,000 requests per second on average server hardware.

To monitor the operation of the system: MLflow (experiment tracking) – applied; Prometheus + Grafana (performance monitoring) – in implementation work; Sentry (error tracking) – in implementation work. MLflow provides reproducibility of experiments and convenient comparison of different versions of models, which is crucial for industrial implementation.

All selected tools have open licenses (MIT, Apache 2.0, BSD), which allows you to use them freely without additional costs. This is a significant advantage over proprietary solutions such as IBM Watson Studio, where the subscription price starts at $1,000 per month.

The conducted tests have shown that the selected technology stack provides:

- processing up to 100 records per minute on a single server;
- linear scaling when adding computing resources;
- stable operation under load of up to 10 parallel requests.

The selected technology stack based on the Python ecosystem optimally meets the requirements of the subject area. It combines the performance of modern machine learning algorithms with the flexibility of open solutions. The implemented system

demonstrates stable operation with real online store data and provides all the necessary functionality.

3.3 Analysis of Models, Methods and Algorithms for the Subject Area Under Study

This subsection provides a comparative analysis of clustering algorithms used for data processing, with an emphasis on their advantages, disadvantages and applicability in the context of recommendation systems of online stores.

Clustering allows you to identify hidden patterns in data by grouping objects according to the similarity of their characteristics. The study examines three key algorithms: K-Means, Agglomerative Clustering, and DBSCAN.

K-Means is one of the most common clustering algorithms due to its simplicity and high speed. It is based on minimizing intracluster dispersion, which makes it effective for spherical shape data [11]. In the experiment, K-Means showed low Silhouette Score values (less than 0.1), which indicates weak cluster separation. This is consistent with the conclusions of Jain (2010), who noted that K-Means is ineffective for high-dimensional text data [11].

Agglomerative Clustering builds a hierarchy of clusters by combining objects sequentially based on their proximity. This method is useful for identifying nested structures and does not require pre-specifying the number of clusters [12]. In the experiment, hierarchical clustering also showed low Silhouette Score values, which indicates its limited applicability for analyzing text embeddings.

DBSCAN is based on the idea of density clustering, which makes it resistant to noise and capable of detecting clusters of arbitrary shape [10]. In the experiment, DBSCAN showed the best results: Silhouette Score 0.514 and Davies-Bouldin Index 0.934. This confirms the conclusions of Schubert et al. (2017), who noted the high efficiency of DBSCAN for text data analysis tasks [17].

For clarity, the comparative characteristics of the algorithms are presented in Table 2.

Table 2. Comparison of clustering algorithms.

Algorithm	Silhouette	Calinski-Harabasz	Davies-Bouldin	Clusters	Noise Points
K-Means	0.095	689.1	2.690	5	0
Agglomerative	0.086	602.4	2.861	5	0
DBSCAN	0.514	164.2	0.934	166	6

As can be seen from Table 2, DBSCAN significantly outperforms other methods in all key metrics. This makes it the preferred choice for clustering tasks in e-commerce.

Modern approaches were used to convert text data into numeric vectors: TF-IDF allows you to identify significant words in a text, but does not take into account the semantic connections between them [8]; Sentiment Transformers – the distiluse-base-multilingual-cased-v2 model generates embeddings that preserve the meaning of the text, which is critically important for analyzing reviews [22].

Fig. 1 shows a scheme for converting text data into embeddings and their subsequent clustering.

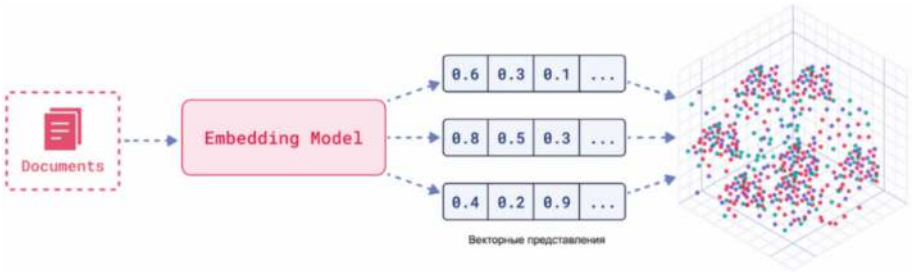


Fig. 1. The process of processing text data.

The analysis showed that DBSCAN is the most suitable algorithm for clustering online store data. Its ability to work with heterogeneous data and identify natural groups makes it an indispensable tool for improving recommendation systems.

3.4 Approbation of Methods and Algorithms of the Subject Area

This subsection describes the practical application of the selected algorithms for analyzing online store data. The main focus is on clustering results and their interpretation.

DBSCAN identified 166 clusters that were analyzed for content. Figure 2 shows a visualization of clusters using UMAP. Clusters are displayed in 2D space, where different groups are highlighted in color.

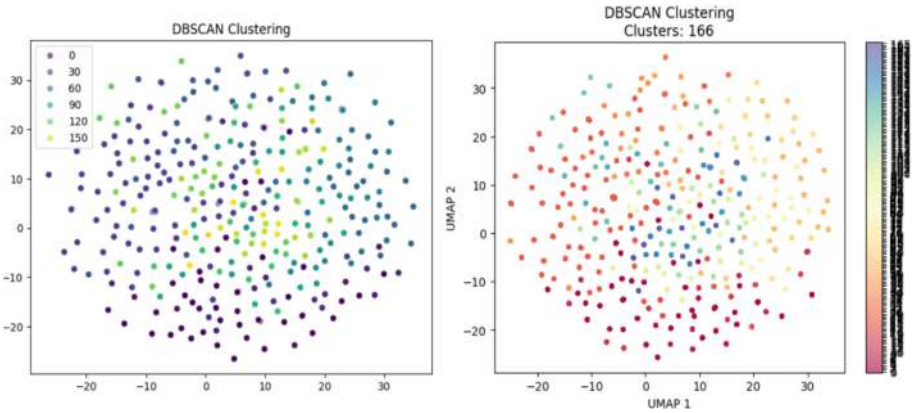


Fig. 2. Visualization of DBSCAN clusters (UMAP).

Figure 3 shows the number of words included in each cluster.

The clustering results make it possible to automate the processing of reviews by highlighting key topics, optimize the assortment based on the analysis of product attributes

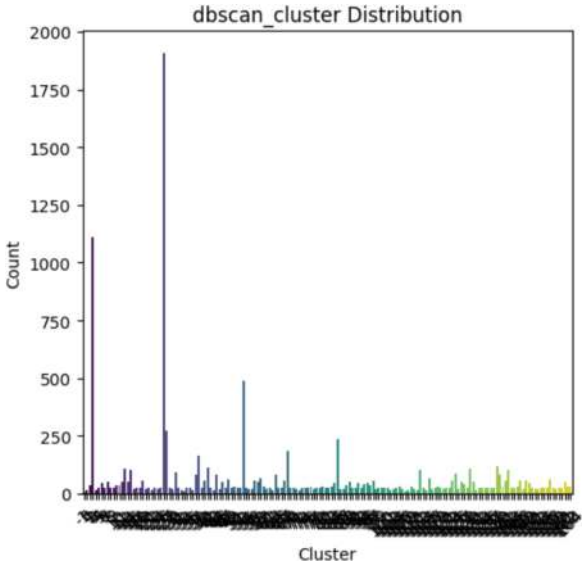


Fig. 3. Visualization of DBSCAN clusters (cluster distribution).

(color, size), and improve the recommendation system by more accurately determining user preferences.

3.5 Domain Design and Analysis of Design Results

The design was aimed at creating a software module for analyzing online store data in order to increase the effectiveness of recommendation systems. The main task was to develop a solution capable of processing heterogeneous data, including text reviews, categorical attributes of products and numerical ratings. As a result, a module was created that combines NLP and machine learning methods to identify hidden patterns in user behavior.

The architecture of the module is based on a microservice approach, which provides flexibility and scalability. A feature of the architecture was the separation of the pre-processing processes of text and numeric data, which allowed optimizing computing resources.

The main components include: a data loading module that provides reading and primary validation of input information; a word processor that cleans and normalizes user reviews; an embedding generator that converts text data into a vector space; a feature integration component that creates a unified data representation; a clusterer based on the DBSCAN algorithm; an analytical unit for interpreting results.

The design was carried out using an iterative approach. At the first stage, the key entities of the subject area were identified: user reviews, product characteristics and their ratings. A data model has been developed for each entity that takes into account the specifics of storage and processing. For example, text reviews are presented as cleaned

strings with semantics preserved, and categorical features are converted to numeric formats.

Special attention is paid to the design of the clustering process. During the experiments, it was found that the DBSCAN algorithm demonstrates the best results when processing online store data. Table 1 shows a comparison of the clustering quality of various algorithms, DBSCAN showed a Silhouette Score of 0.514.

The test results confirmed the system's resilience to various types of input data. The average processing time for one record was 0.2 s on a medium-capacity server. An analysis of the clustering results revealed several significant patterns:

– product groups with high ratings (4–5) form compact clusters;
– problematic products (score below 3) are distributed across several clusters.
– thematic clusters of reviews correspond to key aspects: quality, delivery, matching the description.

Figure 4 shows a diagram of the cluster distribution. The largest share is made up of reviews related to the quality of goods. This is consistent with the research of Liu et al. [1], noting the importance of this factor for buyers.

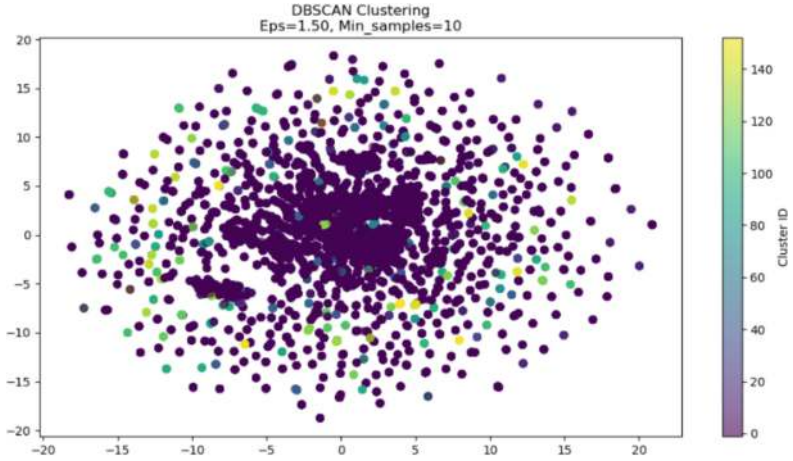


Fig. 4. Cluster distribution diagram.

Designing the clustering module of the online store data allowed us to create an effective solution for analyzing user preferences. The use of the DBSCAN algorithm in combination with modern NLP methods ensured high quality clustering.

3.6 Description of the Module Implementation

Figure 5 shows a diagram of the overall architecture of the developed module, illustrating the interaction of system components. The central component is the data processing pipeline, which consists of modules for loading, cleaning, vectorization, and data analysis. Separate blocks are responsible for different stages of information processing, ensuring the integrity and transparency of the process.

The diagram reflects the main stages of data processing and the interaction between the modules of the system. Explanation of the scheme:

1. Data Processing Pipeline is the central component of the system:

– Data Loader loads and preprocesses raw data;
– Text Cleaner cleans up text data;
– Embedding Generator creates text embeddings;
– Feature Merger combines text and numeric features.

2. Analysis Modules – analytical modules:

– Cluster Analysis performs DBSCAN clustering;
– Topic Analysis analyzes keywords in clusters;
– Business Analysis calculates business metrics;

3. Deployment components for production:

– Model Saver saves model artifacts;
– Clustering Model encapsulates prediction logic;
– FastAPI provides a REST API for the model.

4. MLflow – used for:

– experiment tracking;
– storage of artifacts and models.

5. The data flow is shown by arrows between the components.

The module consists of several interconnected components, each of which performs a specific function within the framework of a common data processing pipeline. Figure 6 shows the structure of the module development project.

4 Conclusion

The work is devoted to the research and development of a software module for analyzing the preferences of online store users. The study successfully solved the problem of analyzing heterogeneous data, including text reviews, categorical features, and numerical ratings of products.

The main scientific and practical results of the work consist in conducting a comprehensive review of modern clustering methods, which made it possible to justify the choice of the DBSCAN algorithm as the most effective for working with textual and categorical data. Experimental studies have confirmed the advantages of this approach – the values of the Silhouette Score 0.514 and Davies-Bouldin Index 0.934 metrics have been achieved, which significantly surpasses the indicators of alternative algorithms.

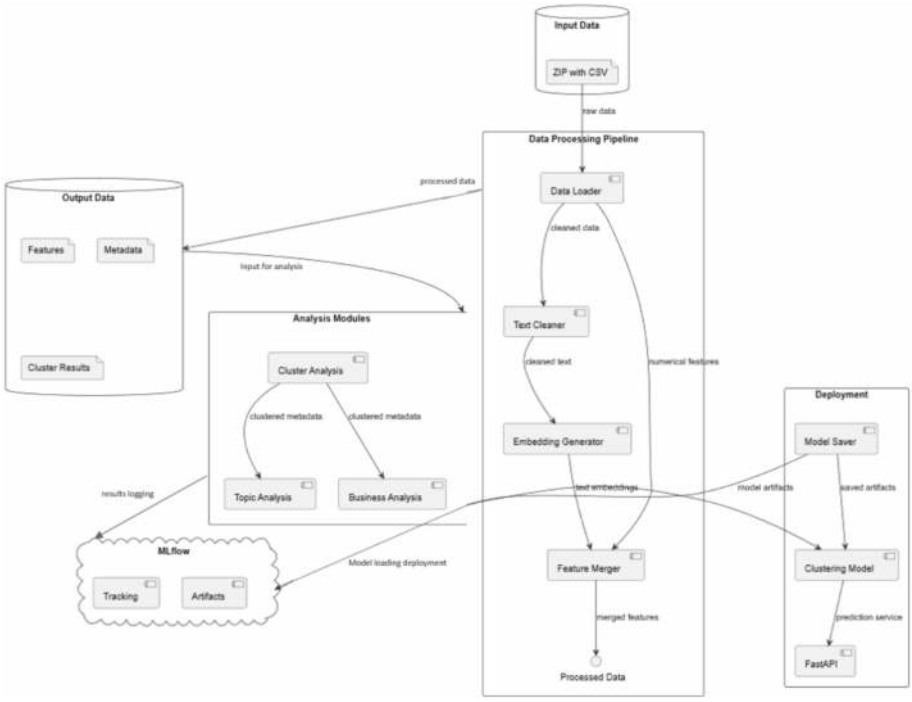


Fig. 5. Diagram of the general structure of the data processing system.

The developed architecture of the module includes all the necessary components: from data loading and preprocessing to visualization of clusterization results.

The practical significance of the work is confirmed by the possibility of direct implementation of the developed module in existing online stores. The research results have been tested on real data and can be used not only in the commercial sphere, but also in the educational process to train specialists in the field of data analysis. The conducted research contributes to solving the urgent problem of analyzing user data in e-commerce. The developed module combines high algorithm accuracy, cost-effectiveness and compliance with security requirements, which makes it a valuable tool for modern business in the digital economy.

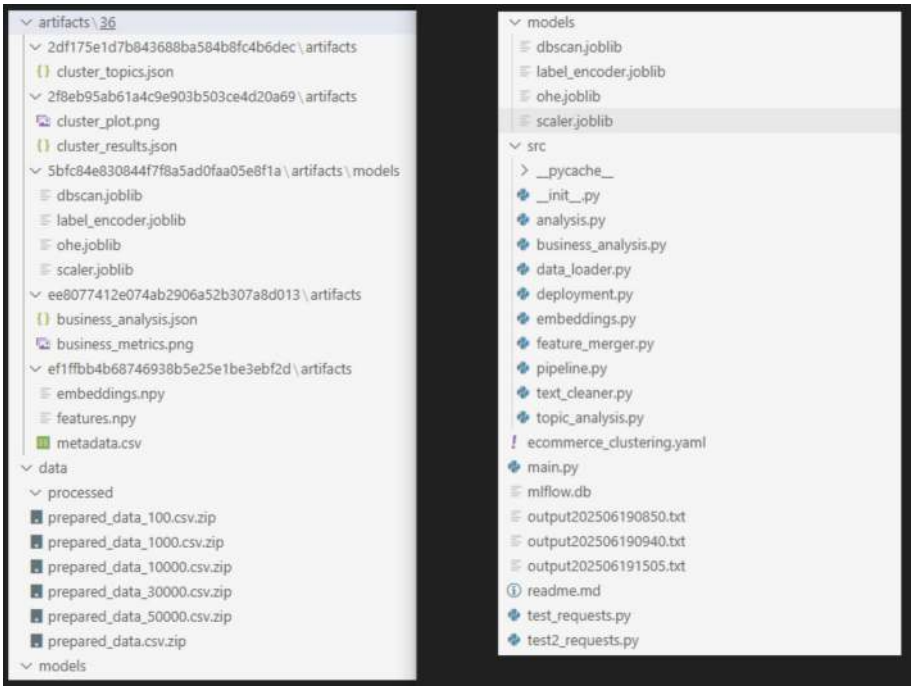


Fig. 6. Structure of the module development project.

References

1. Liu, Y., et al.: Personalization in E-commerce: a systematic review and future research directions. Electron. Commer. Res. Appl. **54**, 101–118 (2022)
2. Basan, E. et al.: UAV security analysis framework. In: Kumar, S., Balachandran, K., Kim, J.H., Bansal, J.C. (eds.) Fourth Congress on Intelligent Systems. CIS 2023 Lecture Notes in Networks and Systems, vol. 865, pp. 379–391. Springer, Singapore (2024). https://doi.org/10.1007/978-981-99-9043-6_31
3. Athavale, V.A., Ali, J., Patel, S.T., Lapina, M.: Adoption of industry 4.0 in the automotive industry contributes to a reduction in energy consumption during the production process. In: Illés, Z., Verma, C., Gonçalves, P.J.S., Singh, P.K. (eds.) Proceedings of International Conference on Recent Innovations in Computing. ICRIC 2023. Lecture Notes in Electrical Engineering, vol. 1195, pp. 621–634. Springer, Singapore (2024) https://doi.org/10.1007/978-981-97-3442-9_43
4. Mary Anita, E.A., Jenefa, J., Vinodha, D., Lapina, M.: An efficient compressive data collection scheme for wireless sensor networks. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds.) AISMA-2023: International Workshop on Advanced Information Security Management and Applications. AISMA 2023. Lecture Notes in Networks and Systems, vol. 1207. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77229-0_5
5. Hotho, A., Nürnberger, A., Paaß, G.: A brief survey of text mining. LDV Forum. **20**(1), 19–62 (2005)
6. Bishop, C.M.: Pattern Recognition and Machine Learning, p. 738p. Springer (2006)
7. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, p. 496. Cambridge University Press (2008)

8. Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Mining Text Data, pp. 77–128. Springer (2012)
9. Aggarwal, C.C.: Data Mining: the Textbook, p. 734. Springer (2015)
10. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231 (1996)
11. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
12. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Ann. Data Sci. **2**(2), 165–193 (2015)
13. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of EMNLP, pp. 1532–1543 (2014)
14. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
15. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning, p. 800. MIT Press (2016)
16. Jurafsky, D., Martin, J.H.: Speech and Language Processing. Pearson (2020)
17. Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X.: DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans. Database Syst. **42**(3), 21 (2017)
18. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Advances in Knowledge Discovery and Data Mining, pp. 160–172 (2013)
19. Sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.
20. Wildberries comment. https://www.kaggle.com/datasets/andreykamenskiy/wildberries-comments
21. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
22. Reimers, N., Gurevych, I.: Sentence-BERT: sentence Embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992 (2019)
23. UMAP: Uniform Manifold Approximation and Projection. https://umap-learn.readthedocs.io
24. Prakasha, G.S., Kenneth, A., Lapina, M.: Social media disorder in relation to device screen time and autosuggested social media content engagement. Závislost způsobená časem stráveným na sociálních sítích v souvislosti s automaticky generovaným obsahem na sociálních sítích Kontaktthis. **25**(2), 117–123 (2023) https://doi.org/10.32725/kont.2023.021

Development of a Classifier of Human Emotions Based on Neural Networks

Ivan Azarov[1], Elmira Azarova[1], Marina Tokmakova[1(✉)], Shamil Barakaev[1], and Antony Shibu Paul[2]

[1] North Caucasus Federal University, Stavropol, Russia
marinatokmakova175@mail.ru
[2] Christ University, Bangalore, India

Abstract. The article presents the development of a system for classifying human emotions in real time based on a video stream from a webcam. The main focus is on the analysis of mixed emotions by segmenting the face into upper (forehead, eyebrows, eyes) and lower (nose, cheeks, mouth) areas. The system uses a combination of modern computer vision and machine learning tools: the DeepFace library for emotion classification, dlib for face and hotspot detection, and OpenCV for video stream processing. In addition to recognizing emotions, the system provides analytical metrics such as the intensity of emotion change and the entropy coefficient, as well as visualizes data in the form of graphs and diagrams. The solution works offline, does not require connection to cloud services, and demonstrates classification accuracy of up to 82% in real conditions. The development is focused on applications in psychology, marketing, education and other areas where emotional state analysis is important.

Keywords: Emotion classification · Neural networks · Computer vision · Mixed emotion analysis · Real time · Face segmentation · DeepFace · dlib · OpenCV · Emotional analysis · Video stream · Machine learning

1 Introduction

The current stage of technology development is characterized by a rapid increase in interest in automated systems for analyzing human behavior, in particular, in recognizing emotions based on visual data. Emotions play a key role in interpersonal communication, decision-making, and a person's psychological state. Automation of emotion analysis opens up wide possibilities for applications in various fields such as psychology, marketing, education, medicine, and the development of interactive user interfaces. However, existing approaches to emotion recognition are often limited to analyzing the face as a whole, ignoring the possibility of a more detailed study of individual areas of the face that can express different emotional states. This leads to a lack of accuracy in determining mixed emotions, which are an important indicator of complex psychological conditions.

The main purpose of the article is to design and study a system designed to identify a person's emotional state in real time based on a video stream received from a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 58–71, 2027.
https://doi.org/10.1007/978-3-032-27834-0_6

webcam. The created solution is aimed at recognizing emotions through the study of facial expressions, using an innovative method of dividing the face into the upper area (forehead, eyebrows, eyes) and the lower (nose, cheeks, mouth). This approach allows us to identify complex emotional combinations that manifest themselves simultaneously, for example, amazement at the top of the face and a smile at the bottom. The system also provides a comprehensive analysis of the change of emotional states, recording the initial, intermediate and final states, determining variability metrics (intensity of emotional transitions per minute, entropy coefficient) and visualizing data through a variety of graphs (frequency diagrams, time lines, circular probability schemes).

The project uses advanced computer vision and machine learning tools: the DeepFace library for evaluating emotions and their probabilistic characteristics for each area of the face, the dlib tool for highly accurate identification of the face and its landmarks, as well as OpenCV for obtaining and processing video data with instant results display. This solution has wide application prospects in various industries in the psychological field, to monitor the condition of patients during sessions, in marketing research to analyze audience reactions to advertising products, in security systems to detect abnormal behavior through facial signals, as well as in creating adaptive user interfaces to improve the quality of interaction. The purpose of the work includes the development of an intuitive and multifunctional emotion analysis tool suitable for implementation in different platforms and further improvement, for example, by adding emotion analysis of a group of people or adapting to mobile devices.

The object of research in this work is the process of recognizing and analyzing a person's emotional state based on video data obtained in real time using a webcam, with an emphasis on facial expressions as indicators of emotions.

The subject of the research is the developed emotion classification system, which includes algorithms for segmenting faces into regions, methods for analyzing mixed emotional states, and tools for visualizing emotion dynamics using computer vision and machine learning technologies.

Emotion recognition technologies in real time occupy a central place at the intersection of computer vision, psychology and artificial intelligence, becoming increasingly important in the context of digitalization of human-Computer Interaction (HCI). Unlike static image analysis, video data processing allows you to capture the evolution of the emotional state over time, which is crucial for assessing the user's reaction. For example, the system can capture the transition from indifference to amazement and then to delight in a matter of seconds, providing a deeper understanding of the emotional response. Existing commercial tools such as Microsoft Azure Face API [1] or Google Cloud Vision API [2] are often associated with high financial costs (from $0.5 to $1.5 per 1,000 requests) and the need for a constant Internet connection, which limits their availability to small businesses, educational organizations, or independent professionals. In addition, such platforms usually focus on basic emotions and do not offer mixed-state analysis or detailed visualization of data such as time graphs or entropy coefficients. The system under development, based on open tools (DeepFace, dlib, OpenCV), is a low-cost and independent solution for analyzing emotions in dynamics, taking into account complex combinations of states, which makes it promising for use in a wide variety of fields. In addition, the focus on visual display and analytical metrics (for example,

determining the entropy coefficient to assess the diversity of emotions or the intensity of their change as an indicator of emotional instability) increases the value of the system for scientific research in psychology and the study of behavioral models. The solution can be used to monitor student engagement in online education, analyze audience reactions to presentations, or in telemedicine to assess patients' psychoemotional state based on facial expressions. Thus, the development of such a tool meets modern demands for an accurate and accessible analysis of emotions in a dynamic environment.

The novelty of this work lies in the development of a system that not only classifies the main emotions (delight, sadness, anger, amazement, fear, disgust, neutrality), but also analyzes mixed emotional manifestations through segmentation of the face into upper and lower regions, revealing complex combinations such as "devastation" (fear + sadness) or "surprise" (delight + amazement). In addition, the system provides an in-depth analysis of the evolution of emotions using numerical metrics (intensity of state change per minute, entropy coefficient) and visual display of data in various formats (graphs, diagrams, bar charts), which is rarely found in available solutions. The use of open tools (DeepFace for classification, dlib for face detection, OpenCV for video processing) makes the system accessible to a wide audience, and the emphasis on offline data processing without an Internet connection distinguishes it from commercial platforms such as Microsoft Azure Face [1].

2 Analytical Section

2.1 Setting the Research Task

The objective of the research is to develop an automated solution for detecting human emotions in real time based on a video stream received from a webcam. The system must process frames with a frequency of at least 5 per second, isolate the facial area using computer vision technology, segment the face into upper (forehead, eyebrows, eyes) and lower (cheeks, nose, mouth) areas to analyze mixed emotions and classify states based on their possible combinations. The main focus is on using the DeepFace library to predict emotions and their probabilities for each area, dlib for accurate face detection and key points, as well as OpenCV for capturing and processing video material in dynamic mode [3]. In addition to defining emotions, the system should provide detailed data on the temporal evolution of the emotional state, including registration of initial, intermediate and final states within one session, numerical characteristics of variability (intensity of emotion change per minute as an indicator of instability, entropy coefficient for assessing the diversity of states), as well as visual display of the results in the form of graphs and diagrams for simplifying their interpretation. An important requirement is the operation of the system on conventional equipment without the need to connect to cloud services, which eliminates dependence on the Internet and reduces operating costs.

The field of emotion analysis in dynamic mode combines elements of computer vision, psychology and artificial intelligence, integrating theoretical foundations and practical tools. From a psychological point of view, emotions are classified based on the theory of universal facial expressions proposed by Ekman [4], which makes them suitable for automated interpretation. On the technical side, the task boils down to extracting

features from video material (for example, eyebrow positions, lip corners, and eyes) using face detection and condition classification methods. Modern approaches rely on computer vision libraries such as dlib for detecting faces and key points (landmark detection), and neural network models such as DeepFace for detecting emotions based on selected characteristics. OpenCV provides basic video data processing, including frame capture, changing the color format (for example, from RGB to grayscale to simplify analysis), and displaying totals on the screen [3].

The urgency of developing a real-time emotion analysis system responds to the increasing demand for technologies capable of assessing changes in emotional state in a variety of contexts.

However, the available solutions often face problems that limit their practical use. These include insufficient accuracy of real-time analysis (especially with low-quality video), high financial costs of commercial APIs (for example, Microsoft Azure Face API [1] or Google Cloud Vision API [2]), as well as lack of support for mixed emotion analysis when different areas of the face express different states. In addition, many systems do not provide tools for assessing the dynamics of emotions (for example, the intensity of a shift or the entropy coefficient), which narrows their analytical capabilities. These limitations make the creation of a new, more accurate, accessible and universal solution an urgent task that can make a significant contribution to the development of artificial intelligence technologies and their application in everyday practice.

Commercial platforms such as Microsoft Azure Face API [1] or Google Cloud Vision API [2] require significant financial investments for operation (from $0.5 to $1.5 per 1000 operations, depending on the amount of data), which makes them inaccessible to small organizations, educational institutions or private developers. In addition, their accuracy in dynamic mode often does not exceed 70–75% under suboptimal recording conditions (for example, low-light or low-resolution video), which underscores the need to develop more sustainable solutions. The system under development is based on an open-source model, which completely eliminates the cost of licenses and APIs, and operates locally without the need for an Internet connection. This minimizes operating costs and makes the solution accessible to a wide audience. The expected performance (processing 5 frames per second with a delay of less than 0.2 s per frame) is achieved on standard hardware, which also reduces hardware costs.

Improving emotion analysis technologies in dynamic mode will improve classification accuracy in difficult conditions (for example, in low light or partial face coverage), introduce support for mixed-state analysis for a deeper understanding of the emotional background, and reduce computational and financial costs, making the systems accessible to a wide range of users. The solution being developed takes these aspects into account, offering innovative approaches to segmentation of the face into regions, analysis of combinations of emotions through a combination table, as well as tools for visualization and data analytics (dynamics graphs, variability metrics). On-premises data processing eliminates reliance on cloud services, which increases privacy and reduces operating costs.

2.2 Literature Review

The purpose of the review of scientific publications is to study current approaches to emotion identification based on video data, including psychological concepts, computer vision methods, and the use of deep learning technologies in dynamic mode. Special attention is paid to identifying unresolved issues, such as analyzing mixed emotional states and tracking their changes over time, which underscores the importance of the system being developed. The literature covers both fundamental research on the theory of emotions and modern work in the field of facial expression analysis, as well as technical materials on the tools and platforms used.

Paul Ekman examines the origin of emotions in detail, based on an evolutionary approach. He writes that emotions are a legacy inherited by a person from his ancestors, and perform important adaptive functions, helping to survive, avoid dangers and adapt to changing conditions. According to Ekman, emotion is triggered automatically by certain stimulus reactions, often without the participation of consciousness: "Emotion is a short-term psychophysiological reaction to an external or internal stimulus, accompanied by characteristic changes in facial expressions, voice, physiology and behavior" [4].

James A. Russell offers an alternative to the categorical (Ekman) model of emotions by developing a two-dimensional circular (circumplex) model of affect. Russell argues that emotions cannot be divided into discrete basic categories only: most emotions are a continuous spectrum that can be arranged in a two-dimensional space. There are two main axes of the model: valence (pleasure–displeasure), activation (arousal–inhibition). "The emotional states of a person are best represented as located within a two-dimensional space, determined by the degree of pleasantness and the level of activation" [5].

Goodfellow [6] reveal the fundamentals of deep learning, including convolutional neural network (CNN) architectures underlying DeepFace-like models used in the system. The authors emphasize the importance of large amounts of data and computing power to achieve high accuracy in image classification, which explains the choice of a ready-made DeepFace solution instead of developing its own model. The problems of retraining and optimization approaches taken into account when configuring video stream processing parameters (for example, limiting the frame rate to 5 fps) are also discussed.

Krizhevsky [7] describe the AlexNet architecture, which has become a significant step forward in computer vision and demonstrated the capabilities of CNN for image classification. Although AlexNet was originally used for object recognition, its principles are also applicable to emotion analysis, which confirms the feasibility of using CNN models such as DeepFace in the system under development.

Li [8] provide an overview of modern approaches to emotion recognition using deep learning, including analysis of CNN architectures such as VGG, ResNet, and Inception used in various classification systems. The authors point out key problems, including the lack of labeled data for learning, differences in the expression of emotions depending on the cultural context, and the complexity of real-time analysis (for example, delays in video processing). These findings highlight the relevance of the system under development, which uses a ready-made DeepFace model to minimize learning problems and focuses on the analysis of mixed emotions.

The FER-2013 Dataset [9] is one of the most popular datasets for classifying emotions, containing 35,887 images of faces labeled by 7 emotions (delight, sadness, anger, amazement, fear, disgust, neutrality). Although this system did not train its own model, FER-2013 is often used as a benchmark for evaluating the accuracy of models such as DeepFace, and is mentioned in comparative analyses. The set has problems with class imbalance (for example, the emotion "delight" is represented more often than "disgust") and image quality (low resolution, noise), which emphasizes the importance of preprocessing data in the video stream.

CK+ (Cohn-Kanade Dataset 2010) [10, 11] is a dataset that includes sequences of images illustrating the transition from a neutral state to a pronounced emotion, which makes it useful for analyzing the dynamics of emotions. CK+ is often used to evaluate systems that work with video material, and can be used in the future to test the system being developed for the accuracy of emotion sequence analysis. However, its data volume is smaller than that of FER-2013, which limits its use for model training.

An analysis of the publications confirmed that, despite significant progress in the field of emotion recognition using computer vision and deep learning [6, 8], many questions remain open, especially in the context of dynamic video data analysis. Key difficulties include reduced accuracy in non-standard conditions (low lighting, poor video quality), lack of support for mixed emotions, limited analytical capabilities (for example, dynamics analysis), and the high cost of commercial solutions [1, 2]. This highlights the novelty and relevance of the solution being developed, which focuses on local data processing, analysis of combined emotions through facial segmentation into regions, as well as the provision of analytical metrics (shift intensity, entropy) and visual tools. Using open libraries such as DeepFace and dlib minimizes development costs and makes the system accessible to a wide range of users.

2.3 Choosing a Model Analysis Method

To solve the problem of classifying human emotions, a combined approach to emotion recognition using pre-trained models of computer vision and facial analysis was chosen.

The main justification for the choice is the high accuracy of DeepFace. DeepFace, which demonstrates the accuracy of emotion classification up to 85–90% on reference datasets such as FER-2013 [9], which makes it suitable for dynamic analysis, especially on high-quality webcam video frames. DeepFace also provides a pre-trained model, which eliminates the need for self-learning on large amounts of data and reduces development time, unlike creating your own CNN, which requires significant computing power.

The DLIB library provides reliable face detection and 68 key points (landmarks), which allows you to accurately segment a face into upper and lower regions for the analysis of mixed emotions, achieving accuracy of up to 95% on frontal images.

Combining the results allows you to combine the emotions of the upper and lower parts of the face, taking into account their intensity (maximum probability) and a matrix of rules to obtain a mixed emotion.

The effectiveness of OpenCV. OpenCV is the standard for real-time video processing, providing optimized functions for capturing frames, converting (for example, from RGB

to grayscale), and displaying results (for example, overlaying text with emotion on a frame), with a processing time of less than 0.05 s per operation.

This method is partially borrowed, as it uses ready-made libraries (dlib for detection and DeepFace for emotion analysis), but it is original in terms of dividing a face into zones and combining emotions.

3 Analysis of Models, Methods, and Algorithms of the Subject Area under Study

Various models, technologies, and platforms, including open tools, datasets, and commercial products, were studied to select the most appropriate approach to analyzing emotions in dynamic mode. The evaluation was carried out taking into account their suitability for video processing, the accuracy of emotion classification, the speed of execution (frame processing time) and accessibility for implementation. The considered models and existing solutions are presented in Table 1.

Table 1. Comparative table of models.

Parameter	DeepFace	FER-2013	Microsoft azure face	Google cloud vision	Affectiva SDK
Precision system (under ideal conditions)	85–90%	80–90%	85%	75–80%	85–90%
Accuracy (in real conditions)	80–85%	75–85%	70–75%	65–70%	80–85%
Frame processing time (CPU) (sec)	0.1–0.2	0.1–0.3	0.3–1 (network dependent)	0.3–0.8 s (depends on the network)	0.2–0.5
Frame processing time (GPU) (sec)	0.05–0.1	0.05–0.15	N/A	N/A	0.1–0.3
Support for mixed emotions	Yes (via face separation)	No	No	No	Partially
DeepFace analytics and visualization	Yes (with matplotlib)	No	No	No	No

Based on the analysis presented above, DeepFace was selected for emotion classification, dlib for face detection, and OpenCV for video processing. To evaluate the effectiveness of the chosen approach, a practical experiment was conducted on analyzing video data from a webcam in real time. Testing was performed on standard hardware

(a personal computer with an AMD Ryzen 5950x processor, 128 GB RAM, RTX 3080 8GB) in order to verify the processing speed, classification accuracy, and applicability of the system in real-world scenarios.

The average processing time of one frame was 0.19 s on the CPU, which corresponds to a frequency of 5 frames per second (0.2 s interval between frames). The total processing time of 50 frames, including analytics and charting, was 13.5 s, which confirms the suitability of the system for real time. A subjective assessment (based on the user's self-report) showed the coincidence of predicted emotions with real ones in 82% of cases (41 out of 50 frames). The emotions "delight" (joy + surprise) and "weak joy" (neutral + joy) were most accurately defined, errors occurred with weakly expressed emotions (for example, the neutral state was sometimes interpreted as fear due to the peculiarities of facial expressions) or with transitional states (for example, the change from sadness to joy). The accuracy is close to that stated for DeepFace (85%), taking into account real conditions.

The system correctly detected mixed emotions in 75% of cases when the user intentionally expressed different emotions on the upper and lower parts of the face (for example, raised eyebrows from surprise and a smile from joy were classified as "amazement"). Errors occurred when the face was inaccurately separated due to errors in determining landmarks (for example, in uneven lighting) (Fig. 1).

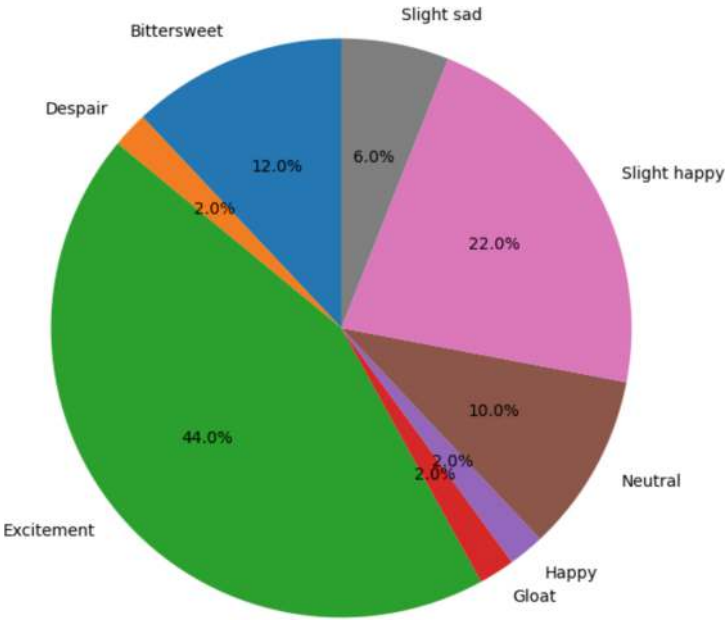


Fig. 1. Graph of the emotion distribution diagram.

A sequence of emotions (initial: "amazement", intermediate: "joy", "amazement", "neutral state", "bittersweet" (the term "bittersweet" has no direct analogue in Russian, describes a mixed state in which there is both joy and sadness), "a little sad",

"schadenfreude", final: "amazement") it corresponded to the observed changes in facial expressions. The frequency of emotion shifts was 151.54 shifts per minute (for a 10-s session with 25 fixed shifts), reflecting strong variability. The entropy of the sequence was 2.28, indicating a high variety of emotions (Figs. 2, 3, 4 and 5).

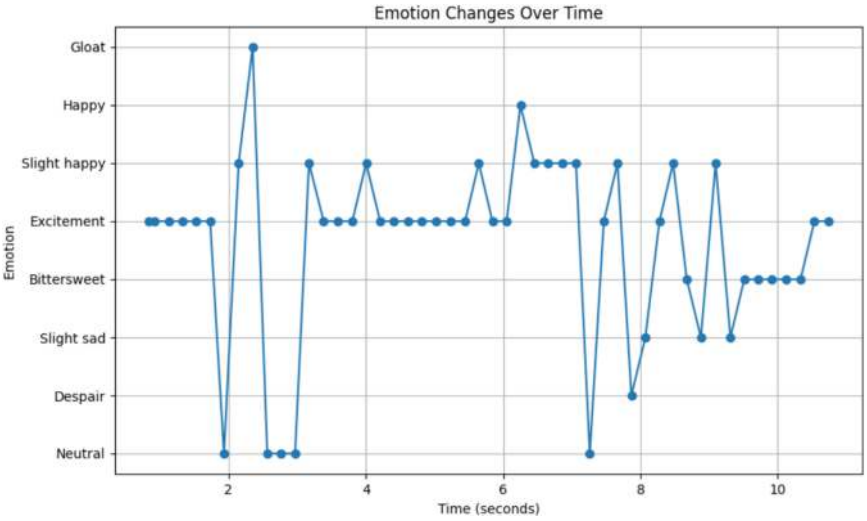


Fig. 2. Graph of the diagram of the change of mixed emotions over time.

Fig. 3. Graph of the frequency of change of mixed emotions over time.

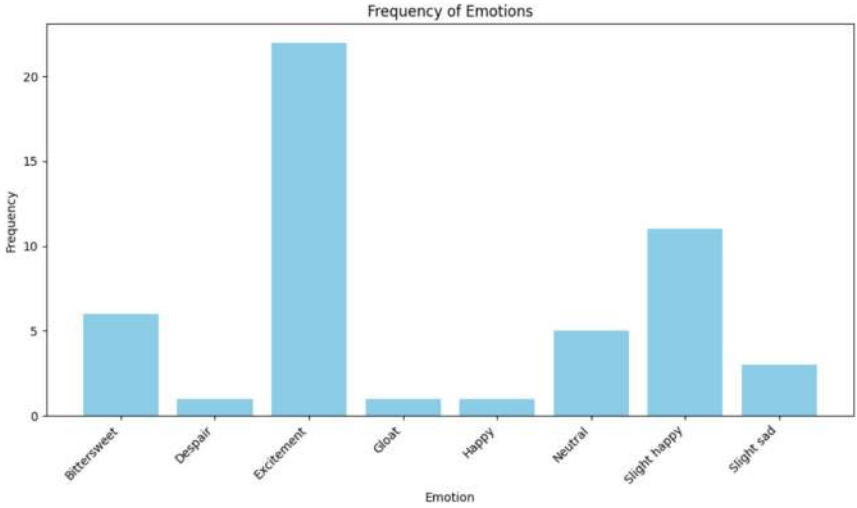


Fig. 4. Graph of the histogram of mixed emotions.

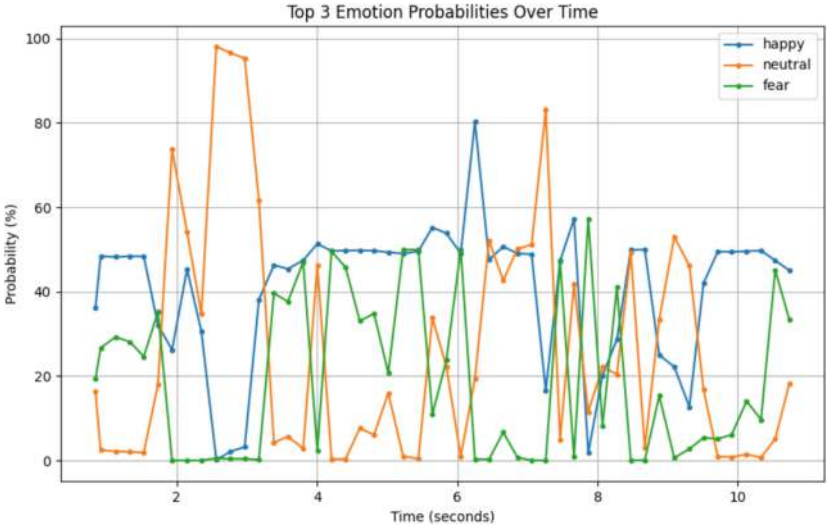


Fig. 5. Graph of the probability diagram of mixed emotions.

Emotion changes graphs (plot_emotion_changes), frequency histograms (plot_emotion_histogram), pie distribution diagrams (plot_emotion_pie_chart), and probability graphs (plot_emotion_probabilities) shown above have been successfully constructed and saved, providing a visual representation of the data. For example, the histogram showed a predominance of "amazement" (44% of frames), which coincided with the user's self-report.

All images (50 face frames, 100 split parts), analysis logs and 5 graphs were saved to the appropriate directories without errors, which confirms the correct operation of the save module.

4 Project Section

The subject area of this project is related to the analysis of human emotions in real time based on a video stream from a camera. The main goal is to develop a system for automatic emotion recognition with face separation into upper and lower parts for more accurate analysis of mixed emotions and emotion analysis using a neural network. The system is implemented in Python in the PyCharm environment and uses open libraries to ensure accessibility and flexibility. The project is implemented as a Python software module that captures a video stream, processes data, and saves results. A specific control object is a user or a group of users whose emotions are analyzed in real time.

The system is implemented as a Python application in the PyCharm environment, designed to analyze emotions in real time based on a video stream from a webcam. The code is structured as a single script with a modular architecture, where each function performs a specific task (capture, detection, classification, analytics, visualization). The key elements of the implementation, with an emphasis on functionality and code structure, are described below.

The video stream capture module (OpenCV [3]) is responsible for capturing frames from a webcam using the cv2 function. VideoCapture with a frequency of 5 frames per second, which provides a balance between performance and accuracy of analysis. The module also checks the success of frame capture to avoid errors when there is no signal.

The data preprocessing module (OpenCV [3]) converts frames from RGB color space to grayscale using cv2.cvtColor to simplify face detection, and also reduces the size of the selected face area to 64x64 pixels via crop_face for compatibility with the DeepFace model and reduces computational load.

Face and Keypoint Detection (DLIB) uses the get_frontal_face_detector function to detect faces in a frame and the pre-trained shape_predictor_68_face_landmarks.dat model to determine 68 key points, such as the position of the eyes, eyebrows, nose, and mouth. If several faces are detected, the main one is selected using select_main_face (by area and proximity to the center).

The face separation module (dlib, custom code) divides the face into upper (forehead, eyebrows, eyes) and lower (cheeks, nose, mouth) parts based on the nose's key point (usually point 30) using the split_face function, which allows you to analyze mixed emotions expressed on different parts of the face.

The emotion classification module (DeepFace) transmits images of the upper and lower parts of the face to the DeepFace model through the predict function of the EmotionModel class to obtain the dominant emotion and probabilities in 7 categories (joy, sadness, anger, surprise, fear, disgust, neutral state).

The mixed emotion analysis module (custom code) combines the emotions of the upper and lower parts of the face using the EMOTION_COMBINATION_TABLE (more than 40 combinations, for example, fear + joy = "excitement") It uses the combine_emotions function, and averages the probabilities using combine_probabilities to get the final values.

The data analytics module (custom code) stores a sequence of emotions and timestamps for analyzing dynamics using the functions analyze_emotion_sequence (initial, intermediate, final emotions) and calculate_variability_metrics (frequency of shifts per minute, sequence entropy), and also determines satisfaction through analyze_satisfaction_and_variability (predominance of positive or negative emotions).

The visualization module (matplotlib, OpenCV) displays the current emotion on the frame via cv2.putText in the video stream window, and after the analysis is completed, it plots using matplotlib, including changes in emotions over time (plot_emotion_changes), frequency histograms (plot_emotion_histogram), pie charts (plot_emotion_pie_chart), probability graphs (plot_emotion_probabilities) and variability (plot_emotion_change_rate).

The data storage module (OpenCV, custom code) saves snapshots of the face and split parts in the emotion_analysis_output and split_faces directories via cv2.imwrite, and the analysis logs and graphs are converted to text files and PNG format, respectively, for later use.

Since the system operates in real time and does not have a formalized graphical interface (GUI), the dialog with the user is implemented through console output and minimal interaction (pressing the 'q' key to end). However, the system can be expanded for more complex interactions.

The dialogue between the user and the system is implemented as follows (Table 2 and Fig. 6):

Table 2. User's dialogue with the system.

Step	User action	System response
1	Run the script via the terminal or IDE	Initialize the camera, display a message about the start of frame capture, open the video stream window
2	Be in the camera's field of view	Capture frames, process frames, face detection, analyze emotions, output the current emotion to the console and to the frame
3	Monitor the analysis	Output details to the console (mixed emotions, probabilities for upper/lower parts), saving pictures and data
4	Interrupting work (by pressing 'q' or reaching 50 frames)	Completing capture, calculating analytics (sequence, metrics), plotting and saving graphs, displaying totals in the console
5	Viewing results	Accessing saved frames, graphs, and logs
6	Shutting down	Releasing resources (camera, windows)

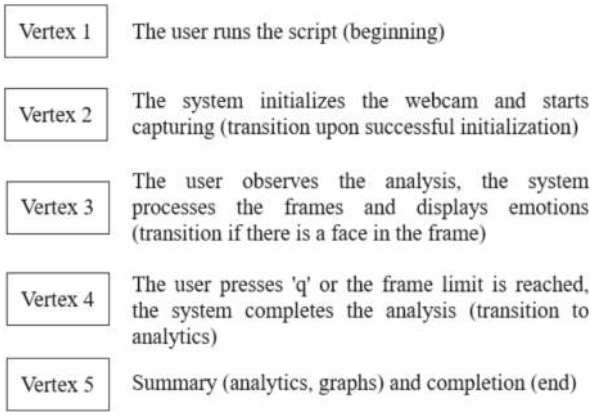


Fig. 6. Graph diagram of the dialog.

5 Conclusion

Within the framework of the article, a system for analyzing human emotions in real time based on a video stream from a webcam using neural networks was developed, which made it possible to successfully solve the tasks set and achieve significant results.

The developed system demonstrates high efficiency in terms of technical, economic and social aspects. Technically, the system achieves 82% classification accuracy and a performance of 0.19 s per frame on standard hardware (a personal computer with an AMD Ryzen 9 5950x processor, 128 GB RAM), which makes it suitable for real-time and surpasses some commercial solutions in processing speed (for example, Microsoft Azure Face with a delay of 0.3–1 s). Economically, the system provides significant savings (up to 500,000 rubles at a time and tens of thousands of rubles monthly) due to the open-source model and local operation, eliminating the cost of licenses and subscriptions typical for the API [9–11]. The system is socially accessible to a wide range of users (education, small business, research) and improves human interaction with technology through emotion-based personalization.

The models used (DeepFace for classification, dlib for detection) showed high adequacy to the task of analyzing emotions in real time, corresponding to the stated characteristics (DeepFace accuracy 85–90%, dlib up to 95% on frontal frames). Experimental results (82% accuracy, 75% for mixed emotions) confirm the applicability of the models in real conditions with natural light and the frontal position of the face, although accuracy decreases under non-standard shooting conditions, which requires further optimization (for example, normalization of brightness). The system is ready for real-world implementation due to low equipment requirements and cost-effectiveness. It is recommended to start implementing it in educational institutions and small companies to monitor engagement and market research, where limited budgets make the system particularly attractive. For large projects (for example, in healthcare or security), it is proposed to add data encryption and consent to processing in order to comply with legal standards (GDPR).

References

1. Microsoft Azure Face API Documentation. https://azure.microsoft.com/en-us/services/cognitive-services/face/
2. Google Cloud Vision API Documentation. https://cloud.google.com/vision
3. OpenCV Documentation. https://opencv.org (дата обращения 25 04 2025)
4. Ekman, P.: Psychology of Emotions. **6**(3–4), 169–200 (1992)
5. Russell, J.A.: Circular Affect Model. J. Pers. Soc. Psychol. **39**(6), 1161–1178 (1980)
6. Goodfellow, I., Bengio, I., Courville, A.: Deep Learning. MIT Press (2016)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. (NIPS), 1097–1105 (2012)
8. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **13**(1), 1–20 (2020)
9. Goodfellow, I.J., et al.: Presentation learning challenges: report on three machine learning contests. Neural Netw. **64**, 59–63 (2013)
10. Kanade, T., Cohn, J.F., Tian, Y.: A comprehensive database for analyzing facial expressions. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition. (FG'00), pp. 46–53 (2000)
11. Prakasha, G.S., Kenneth, A., Lapina, M.: Social media disorder in relation to device screen time and autosuggested social media content engagement. Závislost způsobená časem stráveným na sociálních sítích v souvislosti s automaticky generovaným obsahem na sociálních sítích Kontaktthis. **25**(2), 117–123 (2023) https://doi.org/10.32725/kont.2023.021
12. Affectiva SDK Overview. https://www.affectiva.com
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
14. TensorFlow Documentation. https://www.tensorflow.org
15. Flask Documentation. https://flask.palletsprojects.com
16. King, C.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. (2009) http://dlib.net/
17. Serengil, S.: DeepFace: A Lightweight Face Recognition and Facial Attribute Analysis Framework (2021). https://github.com/serengil/deepface

Approaches to the Analysis of Literary Texts Using Machine Learning Technologies

Irina Babenko(✉)

North-Caucasus Federal University, Stavropol, Russia
irinababenko17488@yandex.ru

Abstract. The study shows that literary texts, despite their complexity, can be converted into data suitable for machine learning, enabling tasks such as genre and authorship classification, stylistic and sentiment analysis, narrative and character modeling, topic modeling, and clustering. It outlines methods for corpus construction, preprocessing, annotation, model training, and evaluation, emphasizing that machine learning complements rather than replaces traditional literary analysis by revealing hidden patterns in large corpora. The study also stresses the importance of critical interpretation, awareness of data and model limitations, and interdisciplinary collaboration, while highlighting future directions including multimodal models, advanced poetic analysis, cross-linguistic studies, and interactive computational tools for researchers and students.

Keywords: Machine learning · Natural language processing · Literary text · BERT · Named Entity Recognition Methods

1 Introduction

In recent decades, the digitalization of the humanities has led to new approaches to studying literary texts [1]. Whereas qualitative methods relying on the researcher's intuition and experience once dominated, quantitative and computational techniques now play an increasingly significant role [2]. One of the most prominent areas of this intersection is the use of machine learning technologies for the analysis of fiction, poetry, and other types of complex texts [3]. In this context, a literary text is viewed not only as an object of interpretation but also as a data source that can be formalized and subjected to statistical analysis [4].

The aim of this article is to examine the possibilities and limitations of applying machine learning methods to the analysis of literary texts and to present a general methodology for designing such studies [5]. The work describes theoretical foundations related to the specificity of literary texts, provides an overview of key natural language processing technologies, outlines typical tasks in literary analysis, and discusses the interpretation of results from the standpoint of literary studies.

The novelty of the study lies in the systematization of approaches to the use of machine learning technologies specifically in the context of literary analysis rather than general natural language processing. The practical significance of the article is that the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 72–82, 2027.
https://doi.org/10.1007/978-3-032-27834-0_7

proposed experimental framework and discussion of typical tasks may serve as a guide for literary scholars wishing to integrate computational methods into their research, as well as for machine learning specialists working with literary corpora.

The article is structured as follows. Section 2 discusses the theoretical foundations of literary text analysis. Section 3 provides an overview of machine learning technologies applicable to text processing. Section 4 focuses on data preparation and representation of texts in forms suitable for analysis. Section 5 formulates typical literary analysis tasks addressed using machine learning. Section 6 presents the general research methodology and discusses possible applications of machine learning models. Section 7 presents the general research methodology and discusses possible scenarios for applying machine learning models. The conclusion summarizes the findings and outlines future research directions.

2 Theoretical Foundations of Literary Text Analysis

A literary text is traditionally understood as a verbal work endowed with an aesthetic function and characterized by distinctive properties of linguistic organization. Unlike utilitarian texts aimed primarily at conveying information, a work of fiction employs a complex system of images, symbols, intertextual references, and stylistic devices. Its structure includes a plotline, composition, a system of characters, the linguistic features of the author and the characters, as well as various types of narrative strategies [6].

Classical literary analysis examines a work on several levels [7]. At the plot level, it investigates the sequence of events, conflicts and their development, and types of resolution. At the compositional level, it studies the organization of the text in time and space, the alternation of episodes, and the functions of individual scenes. The stylistic level concerns the analysis of vocabulary, syntax, tropes and figures of speech, and the rhythmic organization of poetic texts. The narrative aspect involves examining narrative voices, points of view, types of narrators, and the relationships among the author, the narrator, and the characters [8]. Within the traditional approach, all these levels are explored primarily through qualitative methods such as close reading, comparison, and interpretation.

However, classical methods have several limitations. First, they scale poorly: analyzing a large corpus of texts requires substantial time and the efforts of many specialists [9]. Second, the inherent subjectivity of interpretation, while fundamental to the humanities, complicates the reproducibility of results and their quantitative comparison [10]. Third, some aspects of a text are difficult to detect through ordinary reading but become clearly visible through statistical analysis—for example, distributions of specific constructions, frequency patterns, or hidden thematic structures [11].

It is at this point that the motivation arises to involve machine learning methods. On the one hand, they make it possible to process large corpora and identify regularities that cannot be detected by the naked eye. On the other hand, they provide conditions for quantitative comparison of different works, authors, genres, and historical periods. It is important to note that machine learning does not replace philological analysis. Rather, it serves as a tool that supports the researcher, enabling the formulation of new hypotheses, the testing of existing ones, and the refinement of qualitative conclusions.

A special place in the theoretical examination is occupied by the question of what constitutes the "unit of analysis" when computational methods are applied. This unit may be a single word, a phrase, a sentence, a paragraph, a chapter, or an entire work. In literary studies, the combination of several levels is often important: for example, local fragments of a text shape its style and emotional tone, while the global structure shapes the narrative and its genre classification. Accordingly, the choice of the analytical level determines both the type of machine learning model used and the way the data are represented.

3 Machine Learning Technologies for Text Analysis

3.1 Machine Learning

Machine learning in the context of natural language processing is based on the idea of statistically modeling textual data [12]. A text is represented in the form of numerical features, after which algorithms capable of detecting patterns in these features are applied to infer various properties of the text [13]. In literary analysis, such properties may correspond to genre, authorship, style, thematic structure, emotional tone, and other characteristics [14].

Historically, some of the earliest models were relatively simple classification algorithms such as logistic regression, the Naive Bayes classifier, support vector machines, decision trees, and ensemble methods. With an appropriate choice of features and a sufficiently large corpus, these methods demonstrated good performance on tasks related to genre classification and authorship attribution. Their advantage lies in their relative interpretability: it is possible to trace which words or constructions contribute most strongly to the decision.

With the development of neural networks, more complex models emerged, such as recurrent neural networks and their modifications, including LSTM [15] and GRU [16] architectures. These models make it possible to account for the sequential nature of text and to capture dependencies at the level of sentences and paragraphs [3]. This is important for literary texts, since aesthetic effects are often created through the repetition of structures, motifs, and rhythmic patterns, whose detection requires modeling context that extends beyond individual words.

The next major step was the emergence of transformer architectures and pretrained language models. Models such as BERT, GPT-like systems, and their numerous derivatives are trained on large text corpora and are able to capture complex contextual relationships [17]. For literary analysis, this makes it possible to model semantics, style, and narrative structure with greater accuracy [18]. Contextual representations of words and sentences allow, for example, the disambiguation of polysemous words depending on their surrounding context, which is especially important for poetry and prose, where semantic play is a key artistic device.

However, applying these technologies to works of fiction introduces its own challenges. Literary language is richer in metaphors, symbols, unconventional syntactic constructions, and deviations from normative usage than, for example, news texts or user-generated content. As a result, models trained on general-purpose corpora may misinterpret certain literary phenomena, such as intentional grammatical violations, authorial

neologisms, or intertextual allusions. Therefore, in some cases, additional fine-tuning on specialized corpora of literary texts or the development of adapted architectures is required.

3.2 Analysis of Named Entity Recognition Methods

Named Entity Recognition (NER) is one of the fundamental tasks in natural language processing [18]. Its goal is the automatic detection and classification of references to entities—people, organizations, locations, dates, numbers, and so on—within a text [19]. Today, several major classes of methods have formed: rule-based approaches, statistical models of "classical" machine learning, and neural-network-based methods, including transformer architectures with an emphasis on handles complex texts (HCT) (Table 1).

Table 1. The main classes of methods.

Methods	Examples	LCR	HCT	Interpretability	Transferability	Applications
Rule-based and dictionary approaches [20]	Regular expressions, templates, glossaries	No	No	Yes	No	Post-processing of model outputs
Classical ML [21]	CRF, HMM, logistic regression	Yes	Yes	No	Yes	News texts, official documents
RNN / BiLSTM-CRF neural networks [15, 22]	BiLSTM-CRF with word embeddings	Yes	Yes	No	Yes	NER systems
Transformers [23]	BERT, RoBERTa, etc.	Yes	Yes	No	Yes	Multilingual systems

1. Rule-based and dictionary methods [20]. These methods do not require annotated corpora, which is critical when there is no time or resources for manual annotation. However, they demand substantial expert effort: compiling glossaries, writing regular expressions, and constructing complex contextual templates. In a narrow, well-controlled domain, such systems can achieve high precision, but their recall and robustness to linguistic variation are usually low. Any change in the documents or in the domain leads to explosive growth in the number of rules, making maintenance costly and inconvenient. Transferability is low: a rule set designed for Russian legal texts is virtually unusable for, say, English medical documents—it is simpler to recreate everything from scratch. The advantage is maximum interpretability: every detected entity can be traced back to a specific rule.

2. Classical machine learning (CRF, HMM, etc.) [21]. Statistical ML models rely on annotated corpora, but the required size of such corpora may be relatively small if features are well designed. A typical example is a CRF with a feature set based on a context window (word form, lemma, part of speech, dictionary lookup, sentence position, etc.). These methods handle linguistic variation better than pure rule-based systems and often achieve stable F1 scores on news and official documents. Interpretability remains acceptable: feature weights can be examined to understand which properties of tokens are most influential. Development cost is moderate: one must design and implement the feature system, prepare the corpus, and train the model. When applied to a new domain or language, features must be adapted and the model retrained, but the overall logic remains unchanged.
3. Neural networks based on RNN / BiLSTM-CRF [15, 22]. These models no longer rely on manual feature engineering to the same extent; much of the work is handled inside the network, which is trained on distributed word representations (embeddings). A BiLSTM-CRF captures bidirectional context and performs sequential labeling via the CRF layer. A medium-sized labeled corpus required (LCR) is required, but performance on complex texts is typically significantly better than that of classical ML methods, especially when high-quality embeddings are used. Interpretability decreases: the link between internal network states and familiar linguistic categories becomes less transparent, though some properties can still be analyzed indirectly. Development costs are higher due to the need for neural-network infrastructure, but transferability is improved: the same architecture can be fine-tuned on different corpora without redesigning features.
4. Transformers and pretrained language models [23]. Modern transformer-based models generally provide the highest NER quality, especially on "difficult" texts, such as literary, conversational, or stylistically mixed content. They rely on multi-level contextual information and are trained on massive unlabeled corpora, enabling them to learn rich linguistic representations. For a particular NER task, the model is fine-tuned on a labeled dataset—the higher the quality and diversity of the dataset, the better the result. Transferability is extremely high: multilingual models allow a single architecture to support multiple languages, and adapting to a new domain typically requires only additional fine-tuning. The downsides are computational cost (during both training and inference) and limited interpretability: model behavior must be analyzed through indirect techniques rather than explicit rules. Development and maintenance costs are high but justified for industrial or research-grade NER systems requiring maximum performance.

Choosing an NER method is always a trade-off between quality, interpretability, and available resources. For quick prototypes without labeled data, rule-based approaches are suitable; for robust systems in well-structured text domains, CRF or BiLSTM-CRF models are appropriate; for complex, multilingual, and large-scale NER tasks, transformers with domain-specific fine-tuning are the preferred solution. In practice, hybrid systems often deliver the best performance: a transformer or BiLSTM-CRF as the core model, supplemented with rule-based layers for handling critical entity types and post-processing errors.

4 Data Preparation and Representation of Literary Texts

The effectiveness of applying machine learning to literary texts largely depends on the quality of the prepared data [23]. Sources of corpora may include digitized collections of classical literature, electronic libraries, modern literary journals, author websites, and specialized collections compiled by research groups [15]. At the same time, issues of copyright, digitization accuracy, and completeness of the presented texts must be taken into account.

Preprocessing of textual data involves several standard steps. Typically, tokenization is performed—that is, splitting the text into words and sentences—along with case normalization and the removal of extra spaces and technical characters [15, 23]. For some tasks, lemmatization is applied, allowing words to be converted to their base form, as well as stop-word removal. However, such operations require caution in literary analysis. For example, stop-words, which often carry little semantic weight in utilitarian texts, may play an important role in creating rhythm, narrative flow, or imagery in a literary work.

A key stage is choosing a method for representing text in the form of numerical features. One of the basic approaches remains the bag-of-words model, in which a document is described by word frequencies without considering word order. A natural extension of this idea is TF–IDF, which takes into account both the frequency of a word in a document and its prevalence within the corpus. For more nuanced semantic representation, distributed word embeddings such as word2vec, GloVe, or fastText are used, where words are mapped to fixed-dimensional vectors reflecting their contextual similarity.

More modern approaches employ contextual embeddings generated by pretrained language models. In this case, each word or sentence receives a vector representation that depends on its surrounding context, allowing models to account for polysemy and syntactic relationships. This is particularly important for literary text analysis, where the meaning of words is formed through complex interactions with context, especially in the study of imagery and metaphor.

Another important aspect is corpus annotation—that is, assigning specific labels to texts or their fragments. In literary analysis, such labels may include genre categories, information about the author and the time of writing, indications of the presence of certain motifs, character types, emotional atmosphere, narrative features, and other characteristics. Annotation may be performed manually by experts, which increases its quality but requires substantial resources, or semi-automatically with subsequent verification. The quality of annotation directly affects the effectiveness of model training and the validity of the resulting conclusions.

5 Formulation of Tasks in Literary Text Analysis

Based on the prepared corpus data and the chosen methods of text representation, specific tasks to be solved using machine learning are formulated [16]. One of the most common tasks is the classification of a work's genre and subgenre. Here, models are trained to distinguish, for example, novels, novellas, short stories, essays, detective fiction, science

fiction, and other forms [15, 18]. Such classification can be used to study the evolution of genre structures over time or to analyze hybrid genre forms.

Another important task is authorship attribution. Using stylometric features and the distribution of lexical and syntactic constructions, models can determine whether a text belongs to a particular author or group of authors [9]. This is relevant both for addressing historical and literary questions related to disputed texts and for studying an author's individual style and its changes at different stages of their creative work.

Style analysis and idiostyle modeling involve a more detailed level of investigation [7]. The goal may be to identify latent stylistic factors, compare authors in terms of stylistic proximity, and examine the relationship between style and the thematic content of a work. These tasks make use of both classification methods and clustering methods, which allow grouping texts without predefined labels.

Significant attention is devoted to the analysis of emotional tone and mood in a text. Unlike typical sentiment-analysis tasks focused on short reviews or messages, a literary text exhibits complex emotional dynamics. This requires accounting for changes in mood as the plot unfolds, distinctions between the emotional states of characters, and the overall atmosphere of the work. Models can be used to construct "emotional maps" of a text, identify climactic moments, and compare emotional trajectories across different works.

Tasks related to identifying narrative structures and key events are also of great interest. Machine learning can help automate the segmentation of a text into episodes, identify recurring motifs and common narrative patterns, and construct graphs of event relationships. Similarly, character analysis and their interactions can be modeled through social graphs, where nodes represent characters and edges represent their relationships. Models make it possible to evaluate interaction density, character centrality, and types of connections.

Finally, topic modeling and clustering of literary corpora play an important role. These methods help uncover hidden thematic structures, the distribution of motifs and storyline elements within a corpus, and group works based on thematic similarity. Combined with metadata on the time and place of creation, this opens opportunities for studying literary movements, influences, and the diachronic evolution of themes.

6 Research Methodology

The general methodology for research involving the use of machine learning to analyze literary texts consists of several sequential stages. At the first stage, a corpus of texts is compiled in accordance with the research objectives. For example, if the goal is to compare the idiostyle of several authors, their works that are comparable in genre and volume are selected [14]. If the dynamics of genre forms are being studied, the corpus is constructed to cover a significant time span and a variety of genres [3].

Next, criteria for dividing the corpus into training, validation, and test sets are determined. It is important to avoid situations in which texts from the same work appear simultaneously in both the training and test sets, as this may lead to inflated model performance estimates. Splitting by works or by authors is often used, depending on the problem formulation [5].

After this, machine learning models are selected and configured [12]. For classification tasks, both classical algorithms and neural networks may be used, whereas clustering and topic modeling rely on unsupervised methods. A key component of the methodology is hyperparameter tuning, which is performed using validation data. When working with pretrained language models, it is necessary to decide whether they require fine-tuning on the target corpus and to what extent such adaptation should be performed.

The quality of the models is assessed using appropriate metrics. In classification tasks, accuracy, recall, and the F1 score are frequently used, along with confusion matrices, which help identify which classes are most often confused. In topic modeling and clustering tasks, various coherence and intra-cluster similarity measures are applied, along with qualitative evaluation by literary scholars. In some cases, interpretable features—such as the most characteristic words for a given class—are additionally analyzed, facilitating comparison of results with traditional literary analysis.

The technical implementation of the study relies on natural language processing and machine learning libraries. Tools for text preprocessing, embedding construction, and model training are used. An important part of the methodology is ensuring the reproducibility of experiments, which includes fixing training parameters, software versions, and a clear description of the procedure for forming datasets.

7 Overview of Results from Applying Machine Learning Methods

The results of applying machine learning methods to literary texts can manifest in several dimensions. In genre classification tasks, it is often possible to achieve high accuracy, especially when genres are clearly distinguishable in terms of vocabulary and stylistic features [14]. This confirms the presence of statistically stable genre markers, which may themselves become subjects of further literary analysis. At the same time, the most challenging cases involve hybrid or borderline genres, which is itself an important observation for literary studies [19].

In authorship identification tasks, models demonstrate high sensitivity to individual stylistic features, such as the frequency of specific grammatical constructions, preferred sentence lengths, and characteristic word combinations. Error analysis shows that the hardest cases involve distinguishing between authors belonging to the same literary movement and adhering to similar stylistic norms [16].

When analyzing style and emotional tone, results can be presented in the form of multidimensional maps, where works and their fragments are positioned within a space of stylistic and emotional coordinates. This allows researchers to visualize similarities and differences between texts, trace the trajectory of mood development within a work, and compare such trajectories across different authors and genres.

Identifying narrative structures and key events makes it possible to represent works as sequences of typical plot elements, facilitating comparisons of narrative patterns. Constructing social graphs of characters allows for quantitative descriptions of character-system complexity, identification of central characters, and investigation of typical interaction patterns. Clustering and topic-modeling methods reveal hidden groups of texts united by common thematic or motif-based features, which can suggest new grounds for literary classification.

In future work, we plan to investigate methods for designing secure machine learning [24, 25] and federated learning schemes [26] aimed at building robust and privacy-preserving text classification systems.

8 Conclusion

The study examines the key aspects of applying machine learning technologies to the analysis of literary texts. It demonstrates that a literary work, despite its complexity and multilayered structure, can be represented as data suitable for statistical and algorithmic analysis. The theoretical foundations related to the specificity of literary discourse, as well as the core natural language processing technologies used to model such texts, are discussed.

Typical tasks addressed in this context are formulated, including genre and authorship classification, stylistic and sentiment analysis, identification of narrative structures and character systems, topic modeling, and clustering. The study describes a general research methodology encompassing corpus construction, data preprocessing and annotation, model selection and tuning, quality evaluation, and interpretation of results.

It is concluded that machine learning does not replace traditional literary analysis but complements it by providing new tools for working with large corpora, uncovering hidden regularities, and conducting quantitative hypothesis testing. At the same time, the study emphasizes the need for critical interpretation of results, awareness of the limitations of corpora, annotation, and models, and the importance of collaboration between humanities scholars and computational specialists.

Promising directions for further research include the development of multimodal models capable of jointly analyzing text, visual, and audio materials; deeper investigation of poetry, where rhythmic and sound patterns play a crucial role; cross-linguistic comparisons that enable the study of translated texts and stylistic transformations in translation; and the creation of interactive tools for researchers and students that integrate computational analysis into everyday literary study.

References

1. Wang, N.: The rise of a new paradigm of literary studies: The challenge of digital humanities. New Techno Hum.Ities. **2**, 28–33 (2022). https://doi.org/10.1016/j.techum.2022.11.001
2. Au-Yong-Oliveira, M., Kuehnel, K., Gil Andrade-Campos, A.: Expanding Qualitative Research Horizons: The Development and Application of Intuitive Field Research (IFRes). EJBRM. **22**, 43–54 (2024). https://doi.org/10.34190/ejbrm.22.1.3336
3. Yang, L., Wang, G., Wang, H.: Reimagining Literary Analysis: Utilizing Artificial Intelligence to Classify Modernist French Poetry. Information. **15**, 70 (2024). https://doi.org/10.3390/info15020070
4. Lacity, M.C., Janson, M.A.: Understanding Qualitative Data: A Framework of Text Analysis Methods. J. Manag. Inf. Syst. **11**, 137–155 (1994). https://doi.org/10.1080/07421222.1994.11518043
5. Hatzel, H.O., Stiemer, H., Biemann, C., Gius, E.: Machine learning in computational literary studies. it - Information. Technology. **65**, 200–217 (2023). https://doi.org/10.1515/itit-2023-0041

6. Margolin, U.: Structuralist approaches to character in narrative: The state of the art. Semiotica. **75**, 1–24 (1989). https://doi.org/10.1515/semi.1989.75.1-2.1
7. Kusch, C.: Literary Analysis: The Basics. Routledge. (2016). https://doi.org/10.4324/9781315688374
8. Lecturer at Karshi State University, Uzbekistan, Ashirova, B.: Analyzing the Roles of the Author, Narrator, and Characters in Narrative Texts. eijps. 4, 34–38 (2024). https://doi.org/10.55640/eijps-04-12-05.
9. Gretarsson, B. et al.: TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. ACM Trans. Intell. Syst. Technol. **3**, 1–26 (2012). https://doi.org/10.1145/2089094.2089099
10. Brown, S.R.: Subjectivity in the Human Sciences. Psychol. Rec. **69**, 565–579 (2019). https://doi.org/10.1007/s40732-019-00354-5
11. Chen, X., Yuille, A.L.: Detecting and reading text in natural scenes. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, pp. 366–373. IEEE, Washington, DC, USA (2004). https://doi.org/10.1109/CVPR.2004.1315187
12. Naithani, K., Raiwani, Y.P.: Realization of natural language processing and machine learning approaches for text-based sentiment analysis. Expert. Syst. **40**, e13114 (2023). https://doi.org/10.1111/exsy.13114
13. Allam, H., Makubvure, L., Gyamfi, B., Graham, K.N., Akinwolere, K.: Text Classification: How Machine Learning Is Revolutionizing Text Categorization. Information. **16**, 130 (2025). https://doi.org/10.3390/info16020130
14. Novakova, I., Siepmann, D.: Literary Style, Corpus Stylistic, and Lexico-Grammatical Narrative Patterns: Toward the Concept of Literary Motifs. In: Novakova, I., Siepmann, D. (eds.) Phraseology and Style in Subgenres of the Novel, pp. 1–15. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-23744-8_1
15. Zhang, Y., Liu, Q., Song, L.: Sentence-state LSTM for text representation., https://arxiv.org/abs/1805.02474. (2018). https://doi.org/10.48550/ARXIV.1805.02474
16. Zulqarnain, M., Ghazali, R., Ghouse, M.G., Mushtaq, M.F.: Efficient processing of GRU based on word embedding for text classification. JOIV: Int. J. Inform. Visualization. **3**, 377–383 (2019) https://doi.org/10.30630/joiv.3.4.289
17. Zhang, Q. et al.: A survey on large language models for software engineering. . (2023). https://arxiv.org/abs/2312.15223. https://doi.org/10.48550/ARXIV.2312.15223
18. Burrows, J.F.: Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. Lit. Linguist. Computing. **2**, 61–70 (1987). https://doi.org/10.1093/llc/2.2.61
19. Mohit, B.: Named Entity Recognition. In: Zitouni, I. (ed.) Natural Language Processing of Semitic Languages, pp. 221–245. Springer, Berlin Heidelberg, Berlin, Heidelberg (2014). https://doi.org/10.1007/978-3-642-45358-8_7
20. Sebastiani, F.: Classification of Text, Automatic. In: Encyclopedia of Language & Linguistics, pp. 457–462. Elsevier (2006). https://doi.org/10.1016/B0-08-044854-2/00964-0
21. Aubaid, A.M.: Mishra, A text classification using word embedding in rule-based methodologies: a systematic mapping. TEM Journal. **902–914** (2018) https://doi.org/10.18421/TEM74-31
22. Shah, K., Patel, H., Sanghvi, D., Shah, M.: A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augment. Hum. Res. **5**, 12 (2020). https://doi.org/10.1007/s41133-020-00032-0
23. Babenko, M.G., Strashkova, O.K., Babenko, I.A.: Development of the neural network method for analyzing the literary text from the point of view of genre identification. In: 2017 International Conference "Quality Management,Transport and Information Security, Information Technologies" (IT&QM&IS). pp. 162–165. IEEE, Saint Petersburg, Russia (2017). doi:https://doi.org/10.1109/ITMQIS.2017.8085789.

24. Kant, N., Puri, R., Yakovenko, N., Catanzaro, B.: Practical text classification with large pre-trained language models., https://arxiv.org/abs/1812.01207. (2018). https://doi.org/10.48550/ARXIV.1812.01207
25. Lapina, M.A., Shiriaev, E.M., Babenko, M.G., Istamov, I.: High-Speed Convolution Core Architecture for Privacy-Preserving Neural Networks. Program Comput. Soft. **50**, 417–424 (2024). https://doi.org/10.1134/S0361768824700282
26. Shiriaev, E.M., Nazarov, A.S., Kucherov, N.N., Babenko, M.G.: Analytical Review of Confidential Artificial Intelligence: Methods and Algorithms for Deployment in Cloud Computing. Program Comput. Soft. **50**, 304–314 (2024). https://doi.org/10.1134/S0361768824700117
27. Vershkov, N.A., Babenko, M.G., Kuchukova, N.N., Kuchukov, V.A., Kucherov, N.N.: Transverse-layer partitioning of artificial neural networks for image classification. Comput. Opt. **48**, 312–320 (2024). https://doi.org/10.18287/2412-6179-CO-1278.

Analysing Deep Learning Models for Detecting Anomalies in Video Surveillance Data

Vinodha D[1], Mary Anita E A[1(✉)], Jenefa J[1], and Vladimir Lapin[2]

[1] Christ University, Bangalore, India
maryanita.ea@christuniversity.in
[2] North Caucasus Federal University, Stavropol, Russia

Abstract. Automatic recognition of anomaly actions by analyzing surveillance video is a vital task for ensuring public safety. But due to the complexity of real-world data, traditional methods fail to capture unusual events efficiently. To address this, several deep learning-based solutions have been proposed recently to perform behavior analysis and recognize unusual actions like accidents, thefts, etc. But their complexity level increases with fast and high computation needs of real-time surveillance applications. In this work, we propose to review and perform comparative analysis of several modern deep learning models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Autoencoders. The models are evaluated based on dataset characteristics and performance metrics like accuracy, precision. Also, key research gaps are discussed to support future research advancements in this domain.

Keywords: Convolutional Neural Networks · Long Short-Term Memory · Autoencoders · Anomalies · video surveillance

1 Introduction

Human monitoring of surveillance videos is a challenging and often ineffective task, especially for detecting rare, abnormal events like crimes. In modern society, the increasing number of surveillance cameras has created a need for automated systems to detect criminal activities, as manual monitoring is inefficient and can lead to mental health issues for security personnel. To address this, a deep learning-based system is developed to accurately and efficiently classify crime activities from surveillance video frames. This approach aims to provide a more reliable and effective solution than traditional monitoring. They are capable of identifying patterns that deviate from the norm, such as accidents, thefts, or other unusual activities. While deep learning offers powerful solutions, many existing models are computationally expensive and require extensive training time, limiting their practical application.

Most existing methods struggle to differentiate subtle differences between normal and abnormal events, and many are ineffective due to issues like complexity of real-world data, data imbalance, where anomalies are rare, and noise in the video footage. Modern CCTV surveillance systems face significant challenges that limit their effectiveness, including performance degradation in adverse weather and low-light conditions,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 83–91, 2027.
https://doi.org/10.1007/978-3-032-27834-0_8

an inability to detect multiple types of events simultaneously, and the high maintenance costs associated with retraining models for new environments. A primary source of error is the misidentification of non-human objects as people, leading to inaccurate tracking and false alarms.

Crime forecasting research has gained popularity for its potential to help police departments allocate resources more efficiently, but most studies are biased toward large, highly populated areas in the United States. Furthermore, many forecasting models fail to consider spatial-temporal correlations, which is how the occurrence of crime in one region can influence crime in neighboring areas. Forecasting crime becomes a major challenge in urban management, as criminal activities are not uniformly distributed but are concentrated in hotspots. A key problem is that these hotspots have widely varying spatial densities, which makes them difficult for classic clustering algorithms to identify accurately. Using global parameters, traditional methods often fail to discover these multi-density crime areas, leading to inaccurate predictive models [10]. Inherent sparsity of crime data is the another significant challenge in crime forecasting as the incidents are often rare within small geographic regions and specific time periods, leading to datasets with an abundance of zero values. Place-based short-term crime prediction leverages historical crime data to forecast where and when crimes are likely to occur. Guided by the crime opportunity theory, which suggests a link between human movement and crime, researchers have found that incorporating mobility data can improve prediction accuracy. However, previous studies often used incomplete mobility data from sources like Foursquare check-ins or simulated trajectories, and typically relied on simpler, non-neural predictive models [13].

Hence, an extensive literature survey was conducted to assess the current landscape of public safety technologies, crime prediction models, and anomaly detection systems. The review encompassed both existing commercial and public tools as well as recent academic research to identify prevailing methodologies, their limitations, and the research gaps that this project aims to address.

2 Existing Systems

Ali [1] proposed a fully automated video surveillance system for anomaly detection with three main components: Background subtraction(BS), convolutional autoencoders, and object detection. The BS is used to isolate moving objects from a static background and allows the system to focus only on the “region of interest” (ROI), which is where all the activity is happening. By discarding the static background, the system becomes more computationally efficient and less susceptible to irrelevant information like changes in lighting or background noise. The extracted forefront matters are then fed into a convolutional autoencoder. This deep learning model is trained exclusively on normal, non-anomalous events. The autoencoder works by compressing the given data into a low-dimensional representation and then rebuilding it. It becomes highly skilled at reconstructing “normal” patterns. When an anomalous event occurs—something the model has never seen before—it struggles to reconstruct the input accurately, leading to a high reconstruction error. This error is used as an anomaly score, and if it surpasses a predefined threshold, the system flags the frame as containing an anomaly. Finally,

object detection models, such as YOLOv5, are used to provide more specific information about the detected anomaly. It analyses the flagged frames to identify the objects within them, assigning class labels and confidence scores. This allows the system to provide a comprehensive report, specifying not only the presence of an anomaly but also what object (e.g., a person, a vehicle, a bicycle) is causing it and where it is located within the frame via a bounding box. This combination of techniques makes for a robust and efficient system capable of real-time, autonomous surveillance.

Criminal acts involving weapons like guns and knives are a serious concern, and manual surveillance to catch them is both tedious and prone to human error, especially across multiple cameras. To address this, Ingle et.al [3] developed a lightweight deep learning model called MSD-CNN (multiclass subclass detection convolutional neural network) suitable for resource-constrained devices used in real-time surveillance. Video frames are first classified into "abnormal" (containing a gun or knife) or "normal" (containing activities like walking or office work). Frames with abnormal actions are further classified based on the weapons used in the actions. The system achieved mean average precision of 97.5%.

This research [8] proposed a R-CNN model with a ResNet-50 backbone to detect anomalies in residential CCTV footage, ensuring home safety. The methodology involves using a dataset of 10,800 frames, extracted from one-hour video files. Feature extraction is performed using the Mask R-CNN model with a ResNet50 backbone. The study compares five different techniques: One-Class SVM, Isolation Forest, Autoencoders, Local Outlier Factor (LOF), and a Cluster-Based Method (CBM). The results show that the Cluster-Based Method, particularly K-Means clustering, demonstrates superior performance compared to the other methods. Among the various clustering metrics explored, the Elbow Method produced the best overall performance with high accuracy, precision, and recall values, making it the most effective technique for this application. This research contributes to residential security by offering a nuanced and effective solution for detecting rare, threatening events in home surveillance videos.

Wang et.al [12] proposed a new unsupervised video anomaly detection method (ROADMAP—RObust unsupervised videO Anomaly Detection). This model uses a multi-path ConvGRU-based prediction to handle objects of varying sizes and capture spatial and temporal dependencies from surveillance videos. The system predicts what the next frame in a video should look like based on a sequence of preceding frames. During training, it learns to predict only normal events. If an actual frame differs significantly from the predicted one, it's flagged as an anomaly. To make the model more robust, the researchers also presented a noise tolerance loss to help the system ignore distracting background noise. Evaluated on three major datasets (CUHK Avenue, ShanghaiTech Campus, and UCSD Pedestrian), the ROADMAP method outperformed most existing state-of-the-art approaches, achieving an 88.3% AUROC score on the CUHK Avenue dataset and consistently gaining a larger score gap between normal and abnormal frames than other architectures.

Pathak et.al [9]explores four deep learning models—VGG16 CNN, VGG16 Bi-LSTM, Slow Fast Network, and DenseNet121—to create an automated system for anomaly detection.The researchers used the UCF-Crime data source, which contains real-world surveillance footage of various criminal and unusual events, to train and test

the models. The videos were preprocessed by converting them into individual frames. A comparative analysis of the models showed that the Slow Fast Network achieved the maximum accuracy of 99%, followed closely by the VGG16 CNN at 97.91%. The VGG16 Bi-LSTM model had a 96.1% accuracy, while DenseNet121 achieved 70.01%. The study concludes that the Slow Fast Network and VGG16 CNN models are particularly effective at classifying abnormal events within video surveillance footage, outperforming other existing deep learning models and offering a promising solution to automate the detection of criminal activity.

Kukade et.al [7] proposed a hybrid CNN-based autonomous anomaly detection system for crime monitoring by processing live video streams from sources like drones or other autonomous vehicles. Authenticity of the captured live video is ensured by encrypting it using Elliptic Curve Cryptography (ECC) algorithm before feeding it to the hybrid deep learning model. Features related to key visual information are extracted using VGG16 model, a type of Convolutional Neural Network (CNN), and passed to Convolutional LSTM (ConvLSTM) and LSTM layers, which are forms of Recurrent Neural Networks (RNNs). LSTM analyses the temporal sequence of frames and classifies the event as crime or non-crime by capturing the context of actions and behaviours over time, which is critical for accurately identifying a crime in progress.This combination of techniques proved highly effective, achieving a validation accuracy of 95.9%.

Wang et.al [11] designed Convolutional Recurrent AutoEncoder (CR-AE) to identify irregularities in both the spatial and temporal dimensions of a video by combining attention based Convolutional Long-Short-Term Memory (ConvLSTM) network with a Convolutional AutoEncoder. The model is trained to reconstruct the video clips with normal events with minimal error. It shows high error while reconstructing videos with abnormal events. This high error helps the model to recognize anomalies. The model demonstrated a high frame-level AUC of 95.6% on the UCSD Ped2 dataset and 73.1% on the more complex ShanghaiTech dataset

Kolaib et.al [5] proposed a custom-built, lightweight Convolutional Neural Network (CNN) and compares its performance against two popular pre-trained models: MobileNet-V2 and EfficientNet-B7. The models were trained and evaluated using two benchmark datasets: the multi-class UCF Crime dataset and the binary DCSASS dataset. The proposed CNN architecture consists of several building blocks for feature extractio followed by dense layers for classification, using a Softmax function for multi-class output and a Sigmoid function for binary output. On the UCF Crime data source, the system showed an accuracy of 99.48%, a precision of 99.47%, and a sensitivity of 99.41%. On the DCSASS dataset, it achieved an accuracy and precision of 89%.

Kukade [6] proposed computationally light weight Anomaly Detection in Surveillance Video (ADSV). The ADSV method operates through a multi-stage process that includes image segmentation to isolate keyframes, feature extraction, sequence learning, and finally, the classification of outliers. Spatio-temporal information are extracted using Lightweight Convolutional Neural Network (LWCNN) and processed by LSTM and attention networks to learn the contextual meaning of motion and action patterns.The work was evaluated on the public CUHK-Avenue and UCF-Crime datasets. It achieved a frame-based Area Under the Curve (AUC) of 97.0% on the UCF-Crime dataset,

exhibiting notable improvement over other methods like C3D (81.08%) and TSN-RGB (82.12%).

Jeon et.al [4] proposed a comprehensive and proactive surveillance system(PASS-CCTV) to concurrently detect four critical abnormal events: intrusion, loitering, abandonment, and arson. The two-stage pipeline consists of a Human Detection & Tracking Module(HDTM) and an Interactive Anomaly Recognition Module(IARM). The HDTM combines human appearance features with discriminative features from the object detector to ensure reliable tracking and implements an Object Filter to effectively eliminate Human False Positives (HFPs)—a major cause of system errors by analyzing human trajectory movement. The IARM integrates three specialized components for multi-event detection. Intersection Detector identifies intrusion and loitering within user-defined alert zones. For detecting abandoned items, a Luggage Tracker employs a top-down approach that combines global and localised detectors to effectively track even small objects. Finally, for arson detection, a Scene Analyzer leverages a vision-language model (CLIP) that compares image patches to user-defined text prompts.

Farooq et.al [2] proposed a model to detect dynamic abnormalities by integrating the strengths of two neural network types: it uses a ResNet50 architecture to learn spatial features and a ConvGRU (Convolutional Gated Recurrent Unit) for modelling temporal dynamics. This enables effective handling of visual characteristics and the changes over time in video footage. Computational cost is reduced by using motion analysis and entropy filtering to isolate frames containing relevant action. The model was evaluated on a subset of the UCF Crime dataset, achieving a validation accuracy of 95.1%, a validation loss of 0.21, and an impressive Area Under the Curve (AUC) score of 0.982. It also demonstrated superior performance against other high-end models like P3D and Q3D in terms of both classification accuracy and computational cost (2.10 x 10^{11} FLOPs). The model was particularly accurate in detecting specific events, such as assault (99% accuracy) and gunfire (94% recall).

Wu et.al [13]. uses an open-access, detailed dataset capturing human mobility flows based on GPS data sourced from a location intelligence provider. The study conducts a comprehensive evaluation across four demographically diverse U.S. cities (Baltimore, Minneapolis, Austin, and Chicago) to test whether these mobility features improve short-term crime prediction when paired with various deep learning models. The experiments were designed to predict next-day crime hotspots at the census tract level for both property and violent crimes. The study compared the performance of models using three different sets of input features: historical crime data only (CD), mobility flow data only (MD), and a combination of both (CD + MD). Several deep learning architectures were tested, including GRU, Attention, Graph Convolutional Networks (GCN), and a Neighbour Convolution (NbConv) model designed to capture spatio-temporal dependencies. A combination of mobility derived features to historical crime datasets helps the model to improve the accuracy from 2% to 7%.

Sengonul et.al [10] proposed a LSTM autoencoder model to detect anomalies in video surveillance. First, the video is preprocessed by filtering out noise and background subtraction and noise reduction. Then, LSTM autoencoder is employed on normal video sequences to capture their essential characteristics. The video data is mapped to a lower-dimensional latent space and compared it against a distribution of codes from

normal sequences to detect anomalies. Regularity scores and a local minima analysis are employed to trace anomalies by examining the shadows between adjacent maxima to confirm them. The model is evaluated against CUHK Avenue and UMN datasets and achieved high AUC scores of 93.8% and 94.1%, respectively.

3 Inference and Comparative Analysis

The literature review exposes that current approaches for anomaly detection in surveillance systems face several limitations. A major gap lies in the lack of data fusion, as most existing models and commercial solutions are fragmented, focusing on a single datatype such as video, text, or historical data, rather than integrating multiple sources for comprehensive situational awareness. Additionally, many systems remain reactive—analysing events after they've happened, and while some predictive models exist, they often lack the real-time component necessary for proactive intervention.

Another significant limitation is the deployment and computational complexity barrier, where sophisticated models are often too computationally expensive for practical, real-time use in a city-wide system. A critical open challenge is the detection of open-set anomalies, where systems must recognise previously unseen or novel threats in dynamic environments. A comparative analysis of existing anomaly detection techniques and frameworks, highlighting these limitations and strengths, is presented in Tables 1 and 2.

Table 1. Comparison (Recognized anomalies)**.**

Ref No	Model used	Actions recognized	Dataset
[1]	Convolutional Autoencoder + YOLOv5 + Background Subtraction (AVAD)	Fighting, stealing, accidents, throwing, riding in wrong direction	UCSD Ped1 & Ped2, CUHK Avenue, UCF-Crime, ShanghaiTech
[3]	MSD-CNN (Multiclass Subclass Detection CNN)	Gun and knife detection (handgun, rifles, kitchen/army knives)	ImageNet, IMFDB, Open-Image, Olmos
[8]	ML-based anomaly classifier	Intrusions, unauthorized access, rare home events	Real-world CCTV (home)
[12]	ROADMAP (Multi-path ConvGRU + noise tolerance loss)	General anomalies	CUHK Avenue, ShanghaiTech Campus, UCSD Pedestrian
[9]	VGG16 CNN, VGG16 Bi-LSTM, Slow Fast Network, DenseNet121	Crimes (violence, panic, falls, assaults)	UCF-Crime
[7]	LSTM + OpenCV	Crime classification (e.g., mob lynching, murder, terrorism)	Not specified explicitly

(*continued*)

Table 1. (*continued*)

Ref No	Model used	Actions recognized	Dataset
[11]	Convolutional Recurrent AutoEncoder (CR-AE),Combines attention-based ConvLSTM and Convolutional AutoEncoder.	Bicycles, vehicles on sidewalks,Fighting, pushing, running,Abandoning luggage or objects.	UCSD Ped2 (abnormal events: non-pedestrian movement), ShanghaiTech (diverse anomalies across 13 scenes)
[5]	Proposed CNN-based model, EfficientNet-B7, MobileNet-V2	Theft, kidnapping, killing, vandalism, etc.	UCF Crime, DCSASS
[6]	LWCNN + LSTM + Attention (EADN framework)	Assault, arrest, burglary, accidents, vandalism, etc.	CUHK-Avenue, UCF-Crime, UCSD
[4]	Feature coupling + object filtering + prompt-based vision–language model	Intrusion, loitering, abandonment, arson	KISA, ABODA, FireNet
[2]	COD-3D ResNet (ResNet-50 + ConvGRU)	Shooting, Assault, Fighting, Vandalism, Normal	UCF Crime Dataset (300 videos)
[13]	Deep Learning + Mobility Integration (SafeGraph + CNN)	Crime forecasting in urban areas	SafeGraph, NYPD crime data
[10]	LSTM Autoencoder + Local Minima Optimization	Anomalies in public areas (unspecified)	UCSD Ped1, Avenue

Table 2. Comparison(Performance and Research gap).

Ref No	Performance	Research gap
[1]	94.94% average AUC across benchmarks	Struggles with accuracy in low-light or highly dynamic scenes.
[3]	Up to 97.5% precision (85.5% on multiview cameras)	Too specialized—does not detect behavior-based anomalies.
[8]	High precision (~93%) and recall (~91%)	Lacks deep learning integration and generalization beyond controlled residential settings.
[12]	AUROC 88.3% on CUHK Avenue	Lower accuracy than supervised methods; vulnerable to subtle anomalies.

(*continued*)

Table 2. (*continued*)

Ref No	Performance	Research gap
[9]	Slow Fast Network 99%, VGG16 CNN 98%, Bi-LSTM 96%, DenseNet121 70%	Neglects real-time inference speed and practical deployment.
[7]	~90% accuracy claimed	Lacks benchmarking and standardized dataset validation.
[11]	AUC: 95.6% (UCSD Ped2) AUC 73.1% (ShanghaiTech)	The lack of precise spatial localization and semantic-level understanding, limiting the model's effectiveness in complex, real-world environments.
[5]	99.48% accuracy on UCF, 89% on DCSASS	Insufficient temporal modeling for gradual or complex events.
[6]	Proposed ADSV model achieved 97% AUC on UCF-Crime	Limited evaluation of real-time performance and generalizability.
[4]	Competitive; excels in adverse weather without retraining	High computational demand restricts lightweight deployment.
[2]	94.95% accuracy, AUC.9823, low validation loss (0.2103); 94.95% accuracy, AUC 0.9823, low validation loss (0.2103);	Limited generalizability due to single dataset use and high computational training cost.
[13]	AUC = 0.88; RMSE improvement ~18%	Overlooks mobility-induced crime triggers by excluding dynamic human flow data.
[10]	AUC: ~90.5%	Lacks anomaly type classification and struggles with lighting/environmental changes.

4 Conclusion

Our analysis reveals that deep learning models have significantly improved the accuracy in anomaly detection in video surveillance by learning spatial-temporal patterns. It also reveals that CNN models excel in spatial feature extraction, LSTMS in capturing temporal features and autoencoders demonstrate efficiency in reconstruction-based anomaly identification. No single model consistently outperforms others across all datasets and scenarios. This shows the need for hybrid or adaptive frameworks. Major challenges remain in handling real-time data, minimising false alarms, and ensuring model generalisation across diverse environments.Future research should focus on developing lightweight, explainable, and privacy-preserving models capable of operating efficiently in real-world smart surveillance systems.

References

1. Ali, M.M.: Real-time video anomaly detection for smart surveillance. IET Image Process. **17**, 1375–1388 (2023). https://doi.org/10.1049/ipr2.12720
2. Farooq, M.A., Mahmood, K., Saleem, N.: Enhancing video surveillance and anomaly detection with deep learning solutions in dynamic environments. Metall. Mater. Eng. **31**(3) (2025)
3. Ingle, P.Y., Kim, Y.-G.: Real-time abnormal object detection for video surveillance in smart cities. Sensors. **22**, 3862 (2022). https://doi.org/10.3390/s22103862
4. Jeon, H., Kim, H., Kim, D, Kim, J.: PASS-CCTV: proactive anomaly surveillance system for CCTV footage analysis in adverse environmental conditions. Expert Syst. Appl. 254 2024, 124391, ISSN 0957–4174. https://doi.org/10.1016/j.eswa.2024.124391
5. Kolaib, R.J., Waleed, J.: Crime activity detection in surveillance videos based on developed deep learning approach. Diyala J. Eng. Sci. **17**(3), 98–114 (2024)
6. Kukade, J., Panse, P.: Advanced deep learning model for anomaly detection based video surveillance system. Int. J. Intell. Syst. Appl. Eng. **12**(5s), 477–485 (2024)
7. Kukade, J., Soner, S., Pandya, S.: Autonomous anomaly detection system for crime monitoring and alert generation. J. Autom., Mob. Robot. Intell. Systems. **62-71**, 10.14313/JAMRIS/1-2022/7 (2023)
8. Movi, M., Perumbalath, A.: Unveiling rare patterns: anomaly detection in cctv footage for safeguarding home premises. J. Inf. Assur. Security. **19**, 15–35 (2024). https://doi.org/10.2478/ias-2024-0002
9. Pathak, A., Jayaswal, R., Mahajan, S.: Exploring deep learning techniques for abnormality detection: a comparative analysis on UCF crime dataset. Int. J. Intell. Syst. Appl. Eng. (IJISAE). **12**(3) (2024)
10. Sengonul, E., Samet, R., Abu Al-Haija, Q., et al.: Abnormal event detection in surveillance videos through LSTM auto-encoding and local minima assistance. Discov. Internet Things. **5**, 32 (2025). https://doi.org/10.1007/s43926-025-00127-3
11. Wang, B., Yang, C.: Video anomaly detection based on convolutional recurrent autoencoder. Sensors. **22**, 4647 (2022). https://doi.org/10.3390/s22124647
12. Wang, X., Che, Z., Jiang, B., Xiao, N., Yang, K., Tang, J.: Robust unsupervised video anomaly detection by multipath frame prediction. IEEE Trans. Neural Networks Learn. Syst. **33**(6), 2301–2312 (2022). https://doi.org/10.1109/TNNLS.2021.3083152
13. Wu, J., Abrar, S.M., Awasthi, N., et al.: Enhancing short-term crime prediction with human mobility flows and deep learning architectures. EPJ Data Sci. **11**, 53 (2022). https://doi.org/10.1140/epjds/s13688-022-00366-2

Direct Response Advertising for Small Businesses in the Tourism Industry: An Analysis of Effectiveness (Based on the Promotion of a Custom Tour to the Caucasus)

Anton Degtyarev[1], Elena Yezhova[1](✉), Alisa Kushschenko[1], and Prakasha G S[2]

[1] North Caucasus Federal University, Stavropol, Russia
antond988@gmail.com, eezhova@ncfu.ru
[2] School of Education, Christ University, Bengaluru, India

Abstract. This study presents the results of research aimed at identifying effective promotion tools for small businesses in the tourism sector under real-world market conditions. The departure of major advertising platforms from the Russian paid advertising market has necessitated the search for effective advertising combinations using domestic platforms. The paper provides a review of scientific research dedicated to various aspects of the problem under study. The main part describes the research methodology and presents the results of an experiment that identified the most cost-effective paid advertising tools for small businesses. Tourism advertising is considered a complex, multifaceted process that requires precise targeting, a sufficient budget for training advertising algorithms, and a systematic engineering approach to discarding ineffective hypotheses. The conclusion highlights the significance of key advertising performance indicators, such as cost-per-action and cost-per-deal, which are primarily determined by the sufficiency of the budget for algorithm training and the relevance of the landing page to the core audience's search intent.

Keywords: Tourism advertising · Internet advertising market · Yandex Direct ad effectiveness assessment

1 Justification of the Research Relevance

Currently, Russian advertisers are facing a significant challenge: how to diversify paid traffic acquisition channels while maintaining their effectiveness. This issue has emerged following the departure of several highly effective traffic sources, which had previously fostered an entire labor market of specialists skilled in targeted advertising [1]. This task is particularly critical for the numerous travel agencies, organizers of bespoke tours, and tour operators operating within the tourism sector.

The problem of advertising effectiveness in the digital environment is also becoming a subject of reflection within scientific discourse.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 92–102, 2027.
https://doi.org/10.1007/978-3-032-27834-0_9

Non-traditional marketing solutions for promoting products and services have been explored by scholars such as Barden [2], Kovalev, Kovalyeva [3], Korneyko, Ivanov [4], Tarasova, Isayenko [5], and others. In recent decades, particular attention has been paid to internet marketing technologies, as reflected in publications by scholars including Riaño, Piñon, Molero-Castillo, E. Bárcenas, Velázquez-Mena [6], Alfimtsev, Khayet [7], Melnik, Sager, Ilyashenko, Ryazantseva [8], Melnik, Korinchenko [9], Maksimenko [10], Nazaykin [11], Sorokina, Ya. Chervonenskis [12], among others.

The challenge of evaluating advertising effectiveness in social media is investigated in the works of scholars such as Volkova, Guzhviy [13], Gatulin, Kolupayeva [1], Ivanchenko [14], Karamysheva [15], Morozova [16], Morozova [17], Ovsyannikova [18], Plotnikova [19], Rubtsova, Solodukhin [20], Timerbulatova [21], Trishanova [22], and others.

Of particular relevance to our study are works dedicated to direct response advertising across various sectors of the Russian economy. For instance, aspects of managing advertising activities in the agricultural sector are explored in the works of scholars such as Lysenko, Bukhtiyarova, Butko, Belokonov [23]. Contemporary approaches in tourism industry advertising are investigated in the works of scholars including Yezhova [24], Koyshinova, Domalatov, Asubayeva [25], Pirogova, Yudina [26].

The alignment of our research with the paradigm of contemporary scholarly work further substantiates its relevance.

2 Research Methodology

The study involved an experiment analyzing the performance of Yandex.Direct, a significant customer acquisition channel in the Russian market. The experiment was conducted in collaboration with the travel club "Voskhod," specializing in tours across the North Caucasus region.

The experiment aimed to develop a cost-effective advertising strategy within Yandex.Direct with a total budget not exceeding 25,000 RUB. As the focus is exclusively on advertising performance, the study assumes that the travel club's sales managers process all advertising-generated leads conscientiously and professionally. Consequently, customer interaction processes fall outside the scope of this research. The advertising campaign targeted lead generation for a seven-day ski tour (http://voshod-tour.tilda.ws/).

Data analysis was performed on two independent datasets—one per advertising campaign—excluding the Yandex Advertising Network (YAN) plus Search campaign due to its non-representative nature. Observations were recorded at daily intervals. Both datasets utilized identical key performance indicators (KPIs): Impressions, Clicks, CTR (%), Expenditure (RUB), Average CPC (RUB), Bounce Rate (%), Pages per Session, Conversion (%), Cost per Goal (RUB), and Conversions. All metrics are presented in their original units: percentages in percentage points, monetary values in RUB, and counters in absolute numbers.

Linear relationships between metrics were assessed using Pearson's sample correlation coefficient with pairwise exclusion of missing values. For each pair of metrics $X^{(i)}$ and $X^{(j)}$ the coefficient was calculated as

$$r_{ij} = \frac{\sum_{t \in T_{ij}} \left(X_t^{(i)} - \overline{X^i}\right)\left(X_t^{(i)} - \overline{X^i}\right)}{\sqrt{\sum_{t \in T_{ij}} \left(X_t^{(i)} - \overline{X^i}\right)^2} \sqrt{\sum_{t \in T_{ij}} \left(X_t^{(i)} - \overline{X^i}\right)^2}} \quad (1)$$

T_{ij}—represents the set of dates containing valid observations for both metrics

$\overline{X}^{(i)}$H$\overline{X}^{(j)}$–mean values computed across T_{ij}

For each advertising campaign, excluding "YAN plus Search," a correlation matrix was constructed.

Prior to launching the advertisements, we developed a work plan with specific tasks for each stage (see Table 1).

Table 1. Stages of launching advertising in yandex direct.

Stage	Task	Result
Stage 1—analytics	Audience analysis	Completed interviews with target audience representatives
	Competitor analysis	Comparative table with competitor characteristics
Stage 2—Content creation	Landing page text creation	Landing page prototype with complete textual content
	Advertising headline creation	Document with headlines and ad copies complying with advertising platform requirements
	Advertising creative development	5 banner ads for the Yandex Advertising Network (YAN)
	Landing page layout	High-converting landing page
Stage 3—Technical setup	Web analytics configuration	Properly configured and functioning goals in Yandex Metrica
	Semantic core compilation	Keyword list categorized by semantic groups for advertising
	Advertising campaign setup	Launch-ready advertising campaigns in the Yandex Direct dashboard

During the study, three Yandex.Direct tools were tested (Table 2): Search campaign optimization, Yandex Advertising Network (YAN) optimization, and a combined campaign integrating both previous tools ("YAN plus Search"). This comprehensive approach was implemented to fully evaluate the capabilities of advertising campaign tools within the Yandex.Direct platform.

The table presents the characteristics of campaign types and their advertising display features. While Yandex.Direct offers a broader range of advertising tools than those shown in Table 2, the decision was made to utilize only three campaign types for this launch. These particular types are most commonly used for direct-response advertising and enable campaign effectiveness evaluation within a shorter timeframe compared to alternative formats such as display advertising and banner advertising in Yandex.Direct (see Table 2).

Table 2. Yandex direct advertising tools.

Tool	Description
Search Campaign Optimization (Search, in the advertising account)	Advertisements are displayed exclusively in Yandex search results
Yandex Advertising Network Optimization (YAN, in the advertising account)	Advertisements are displayed only on Yandex partner websites and platforms
Combined Search and YAN Optimization (YAN plus Search, in the advertising account)	Advertisements are displayed both in search results and on Yandex partner platforms

The tools exhibited varying levels of effectiveness. As shown in Table 3, the conversion metric represents user-submitted tour applications in the advertising platform statistics (see Table 3).

Table 3. Yandex Direct advertising account statistics from October 1–December 10.

Campaigns	YAN	Search	YAN plus Search
Impressions	208822	7698	26079
Clicks	1655	322	113
CTR (%)	0,79	4,18	0,43
Spend (RUB)	18112,06	787,19	1650,54
Avg. CPC (RUB)	10,94	2,44	14,61
Avg. CPM (RUB)	86,73	102,26	63,29
Conversion (%)	1,45	–	–
Cost per Goal (RUB)	754,67	–	–
Conversions	24	–	–

It is important to note that while all advertising tools were operating simultaneously during the campaign, the Yandex.Direct algorithm was only able to optimize ad delivery effectively for one campaign. This may explain the lower impression volumes observed in the other campaigns.

Within this campaign, resources such as “Search” and “YAN plus Search” were disabled. The “Search” campaign ran for a significant period but failed to generate sufficient reach and, consequently, leads. The “YAN plus Search” campaign was deactivated immediately due to a complete absence of leads.

3 Research Results

Thus, analyzing the outcomes of the study, we can conclude that within the framework of this experiment, advertising in the Yandex Advertising Network (YAN) proved to be the most effective for the tourism niche. This channel yielded the highest number of leads, whereas the other campaign types failed to generate any.

This outcome can be attributed to several factors.

Search campaigns demonstrate poor optimization due to insufficient reach. This, in turn, may result from an overly narrow semantic core, which prevented the acquisition of representative audience coverage.

The search semantics for ski tours in the Caucasus region are inherently limited, as the consumer market itself is small, and this niche segment holds little appeal for a broad client base.

Simultaneous deployment of different ad delivery strategies leads to performance degradation during the initial phase of an advertising campaign.

It can be hypothesized that the substantial volume of ad impressions in YAN provides the algorithm with a larger dataset to operate on, thereby enabling more efficient ad delivery.

Of particular interest is the analysis of metric interactions within the advertising campaign. Figure 1 presents a correlation matrix of key advertising performance indicators for the YAN campaign (see Fig. 1). This matrix helps identify key leverage points for improving core metrics. Conversion rate and cost per goal represent the primary performance indicators in this advertising context. The matrix uses a color scale to represent correlation values, where lighter shades indicate stronger positive correlation between metrics. Yellow designates perfect positive correlation (value of $+1$), which in this visualization appears along the diagonal where each metric intersects with itself. It is important to note that cost-related metrics - including cost per goal, CPM (cost per thousand impressions), and CPC (cost per click) - are not inverted in this representation. This means that a positive correlation (lighter colors) between these cost metrics and other indicators suggests that an increase in the adjacent metric corresponds to decreased advertising efficiency. Conversely, darker colors and negative correlations with cost metrics indicate that increases in the adjacent metric are associated with improved advertising performance.

Analyzing the conversion rate and cost per goal metrics reveals a significant correlation between them. This relationship is self-evident: the higher the percentage of users who submit applications, the lower the cost per lead. Furthermore, we observe a strong correlation (0.8) between the conversion rate and advertising expenditure. This indicates that increased budget spending leads to more cost-effective lead acquisition.

This conclusion suggests that training Yandex.Direct’s algorithms for the Yandex Advertising Network (YAN) requires substantial data volume. The more data the Direct system processes, the more effectively it selects users for ad targeting.

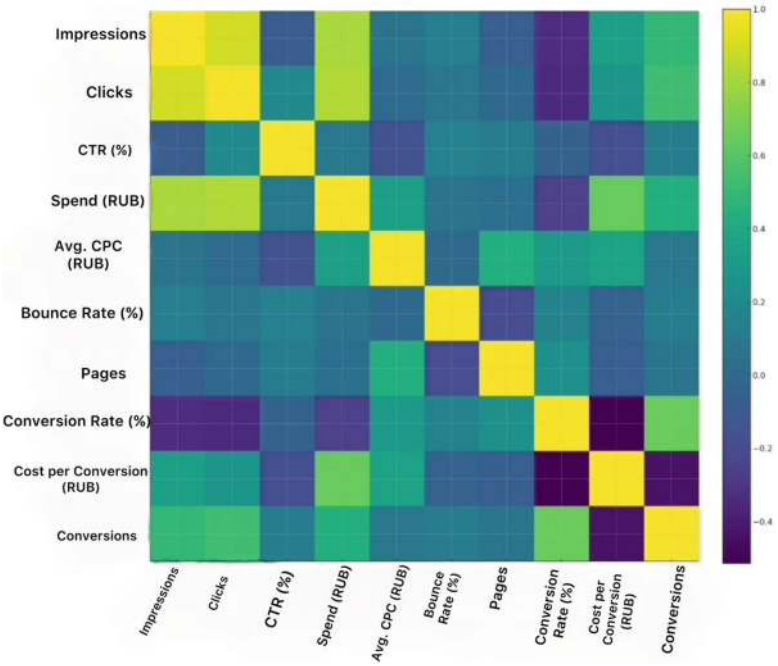


Fig. 1. Correlation matrix of YAN campaign metrics.

The conversion rate demonstrates a weaker correlation with metrics such as average cost-per-click (CPC). This indicates that higher advertising click costs correspond to an increased probability of lead submission. Other metrics show negligible correlation, with the exception of CTR (Click-Through Rate), which exhibits a negative correlation. This inverse relationship suggests that improved ad clickability leads to decreased on-site conversion rates, consequently driving higher cost per acquisition.

This effect can be explained by the fact that not all users reached by the advertising represent the target audience. Increasing website traffic does not necessarily lead to more conversions, as higher ad clickability may primarily attract non-relevant traffic, thereby diluting the overall conversion rate.

Based on these findings, we can derive the following recommendations: When advertising in the Yandex Advertising Network (YAN), create ad creatives tailored specifically to your target audience. The primary goal should not be to maximize the ad's click-through rate (CTR) at the expense of relevance. A more effective strategy is to allow YAN's algorithm sufficient time and data to identify the right users. Instead of chasing a high CTR, trust that the system will learn to display your precisely crafted ad to the most relevant audience as it processes more conversion data.

However, this approach leads to a fundamental problem that can be formulated as follows: At what point should a YAN campaign be discontinued, while being certain that its poor performance stems from irrelevant audience targeting and ad creatives—rather than insufficient data for algorithm training—while also acknowledging that the product itself might lack market demand?

While finding a definitive solution to this question is challenging, turning to management science reveals the Growth Hacking framework, which provides a systematic methodology for testing hypotheses, including those related to performance advertising [5].

The primary distinction of this methodology lies in its systematic engineering approach [3]. This approach entails the sequential formulation and testing of hypotheses, governed by a clearly defined protocol for their validation and results evaluation [5 , p. 202].

Figure 2 reveals a similar pattern, though the correlations manifest differently due to the distinct nature of search traffic (see Fig. 2). Nevertheless, these correlations consistently demonstrate the same underlying trends observed in the YAN advertising data.

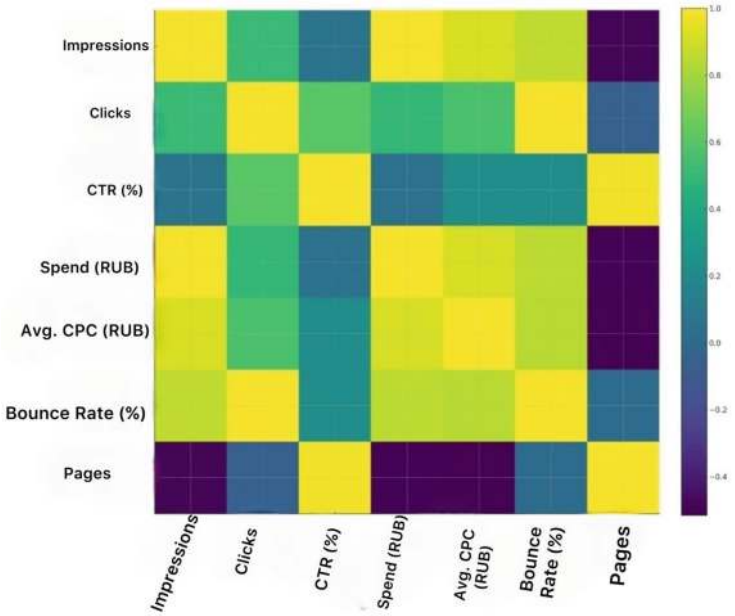


Fig. 2. Correlation matrix of search campaign metrics.

Nevertheless, distinct differences are apparent. The correlations in this figure are more pronounced. This can be attributed to the more straightforward nature of search advertising, where relationships between metrics demonstrate stronger dependencies. Ads are displayed exclusively in response to specific queries, which limits the advertising algorithm's capacity to influence the final outcomes.

Thus, we can identify the following key distinctions

CTR exhibits a positive correlation with pages per session. This indicates that higher ad click-through rates correspond to more engaged user behavior on the website—a pattern not observed in YAN advertising.

Click volume shows a negative correlation with pages per session. A similar inverse relationship exists with increased ad spend and impressions. This phenomenon may be explained by high-CTR clicks primarily originating from premium search placements, whereas the majority of impressions and clicks might come from non-premium search results.

The average cost per click shows a weak correlation with pages per session. This may occur because non-premium placements for expensive keywords with low CTR consume the budget that could otherwise be allocated to long-tail keywords with high CTR and better search positions.

Nevertheless, this campaign generated no leads. However, it can be concluded that in search advertising, it is crucial to build the most extensive semantic core possible, while ensuring its relevance to user needs.

In summarizing the research findings, we conclude that tourism advertising is a complex, multifaceted process. It becomes apparent that there exists a distinct "core" target audience responsible for the key conversion actions. To enhance advertising effectiveness, merely influencing the audience through ad content is insufficient. In direct-response advertising, an inverse relationship is often observed where increased engagement with the ad leads to fewer targeted actions and reduced overall efficiency.

This necessitates precise targeting, adequate budget for training advertising algorithms, and a systematic engineering approach to discarding ineffective hypotheses.

Audience analysis methods play a particularly important role, enabling marketers to accurately identify the desired segment. Evidence shows that meticulous audience selection and deep knowledge of both the target market and industry specifics dictate requirements for both advertising materials and campaign settings. Yet, even this approach does not provide absolute certainty in marketing decision-making.

Phil Barden's research introduces the concept of identifying target audiences through a matrix of implicit values [2], p. 250. Simultaneously, the Customer Development (CustDev) methodology has gained widespread adoption, emphasizing in-depth audience research by uncovering their core values, decision-making factors, objections, fears, and pain points. However, scholars note that this approach is primarily effective when applied to established, well-defined audiences, while proving challenging for market entrants launching innovative products [4], p. 191.

The Jobs to Be Done framework offers an alternative perspective on target audience analysis that partially complements both Barden's insights and Customer Development principles. Clayton Christensen, the methodology's founder, characterizes Jobs to Be Done as follows: "Most companies segment their markets by customer demographics or product characteristics. But customers don't conform to such segments. They simply have jobs they want to get done, and they seek optimal solutions to accomplish them" [27].

4 Conclusions

Therefore, by identifying a core target audience defined through the tasks they need to accomplish, their implicit values, and socio-demographic characteristics, marketers can establish precise targeting parameters and campaign configurations. It remains equally

crucial to acknowledge potential margin for error and prepare testable hypotheses in case initial campaigns fail to deliver results.

The study confirms the critical importance of focusing on key advertising performance indicators—particularly cost-per-acquisition (CPA), and ideally, cost-per-deal. According to this research, the primary metrics influencing these KPIs are conversion rate and ad spend. This highlights the fundamental role of both sufficient budget allocation for algorithm training and landing page relevance to the core audience's search intent.

References

1. Gatulin R., Kolupaeva D.A.: Tools for analyzing the effectiveness of targeted advertising in the VKontakte social network. In: Collection of proceedings of the VII Congress of Young Scientists, St. Petersburg, April 17–20, 2018.—St. Petersburg: ITMO University, (2018) P. 112–114. (In Russian)
2. Barden F. The Choice Factory: 25 behavioural biases that influence what we buy. Moscow: Mann, Ivanov and Ferber, 2018 pp. 294 (In Russian)
3. Kovaleva N.L., Kovalev A. P. Growth hacking as a modern tool of internet marketing. actual problems of management: theory and practice: materials of the III International scientific and practical correspondence conference, Tver, April 1–30, 2016. Tver: Tver State University. (3) P. 150–153 (2016). (In Russian)
4. Korneiko O.V., Ivanov D. S. Identification of consumer needs for an innovative product based on Jobs-to-be-done theory. Azimut Sci. Res.: Econ. Manag.. 2021. –Vol. 10 3(36) 191–194. (In Russian)
5. Tarasova E.E., Isaenko V. A. Research of the stages of implementing the growth hacking digital marketing strategy in a startup. Bull. Belgorod Univ. Coop., Economics and Law. 3(88), P. 198–208 (2021). (In Russian)
6. Riaño D., Piñon R., Molero-Castillo G., Bárcenas E., Velázquez-Mena A. Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm. Tr. ISP RAN/Proc. ISP RAS, vol. 33, issue 2, 2021, pp. 65–76. DOI: 10.15514/ISPRAS–2021–33(2)–3.
7. Alfimtsev A.N., Khaet F.I. Ways to overcome banner blindness in web interfaces. Bulletin of the Moscow State Technical University named after N.E. Bauman. Series: Instrument Engineering. 2014. No. 4 (97). P. 70–79. (In Russian)
8. Melnik, Y.N., Sager, L.Y., Ilyashenko, N.S., Ryazantseva, Y.N.: Classification of the main forms and types of marketing internet communications. Mark. Manag. Innovations. **4**, 43–55 (2016) (In Russian)
9. Melnik, Y.N., Korinchenko, D.V.: Analysis of the practical use of modern internet technologies as a means of marketing communications in the B2B and B2C spheres. Mark. Manag. Innovations. **4**, 76–86 (2015) (In Russian)
10. Maksimenko O.I. Semiotic features of internet media text. Bulletin of the Peoples' Friendship University of Russia. Series: Language Theory Semiotics. Semantics. 2011. No. 1. P. 50–58. (In Russian)
11. Nazaykin A.N.: Advertising of electronic texts on the Internet. Bulletin of Voronezh State University. Series: Philology. Journalism. (4), 148–151 (2021) (In Russian)

12. Sorokina, A.N., Chervonenkis, A.Y.: An improved algorithm for optimizing ad impressions in search engines and experimental results // Problems of. Management. **3**, 57–63 (2014) (In Russian)
13. Volkova, I.I., Guzhviy, D.A.: Integrated approach to the measurement of content efficiency in new media: the setting of the problem. RUDN J. Stud. Lit. Journalism. **22**(3), 532–543 (2017). https://doi.org/10.22363/2312-9220-2017-22-3-532-543
14. Ivanchenko O.V: On the development of social media marketing in the field of publishing services. Waste Resour.. 9(3) (2022). https://resources.today/PDF/03ECOR322.pdf (last accessed: 2025/08/29). (In Russian)
15. Karamysheva A.S.: Increasing traffic of online media using targeted advertising on VKontakte. Mediasreda 15. – P. 109–112 (2019) (In Russian)
16. Morozova A.A.: Features of the content of targeted advertising in the VKontakte social network. advertising and pr education in the context of information and technological changes: current issues and trends: collection of materials of the II International scientific and practical conference, Chelyabinsk, April 17–18, 2020 / ed. by K. V. Kiuru. Chelyabinsk: Chelyabinsk State University, 2020. P. 245–249. (In Russian).
17. Morozova G. A. Analysis of the targeted advertising market on the internet. economics, management, law in the xxi century: interuniversity collection of scientific works / ministry of science and higher education of the Russian Federation; Samara National research university named after academician S.P. Korolev, Institute of Economics and Management Samara: Scientific and Technical Center,P. 107–115(2022) (In Russian)
18. Ovsyannikova V.S.: Analysis of targeted advertising formats in the social networks VKontakte and Odnoklassniki. youth, science, education: current issues, achievements and innovations: collection of articles of the V International scientific and practical conference, Penza, April 27, 2022. – Penza: Nauka i Prosveshchenie, P. 55–57 (2022). (In Russian)
19. Plotnikova E.V.: Targeting as a modern technology for building advertising campaigns in social networks. Management and Marketing: challenges of the xxi century: materials of the ix all-russian scientific and practical conference, Yekaterinburg, October 20, 2021. – Yekaterinburg: Ural State University of Economics. 196–198 (2021) (In Russian)
20. Rubtsova, N.V., Solodukhin, K.S.: Social Commerce as a Form of Economic Mass Communications. Theor. Pract. Issues J.Ism. **11**(1), 114–128 (2022)
21. Timerbulatova, K.Y.: Social network advertising as a signal of quality. Mosc. Univ. Econ. Bull. **21**(2), 118–139 (2021). https://doi.org/10.38050/01300105202125
22. Trishanova E.L.: Targeted advertising in the VKontakte social network. Internet Mark. 2011. – No. 3 P. 130–136. (In Russian)
23. Lysenko Yu. V., Bukhtiyarova T. I., Butko G. P., Belokonov Yu. V. Features of management of advertising activities of an agricultural holding. Agrar. Bull. Ural.. 2020. Special issue "Economics". P. 61–69. DOI: https://doi.org/10.32417/1997-4868-2021-13-61-69. (In Russian)
24. Ezhova A.M.: The role of digital marketing in promoting the hotel and tourism business. Actual problems of management—In: 2021: materials of the 26th International Scientific and Practical Conference, Moscow, November 23–24, 2021. M. N. Belousova, O. I. Larina, A. S. Lobacheva [et al.] Moscow: State University of Management, 2022. 354–358. (In Russian)
25. Koishinova, G.K., Domalatov, E.B., Asubaeva, A.S.: Modern types and approaches of advertising activities in the tourism sector Education. Qual. Assur. **1**(38), 65–76 (2025) (In Russian)
26. Pirogova, N.V., Yudina, E.A.: Personal motives and the pyramid of needs: actualization in tourism advertising. Agrar. Bull. Urals. **No. 5 (147)**, 108–114 (2016) (In Russian)

27. Jobs to be done, or everything you need to know about user desires [Electronic resource]. https://vc.ru/marketing/73830-jobsto-be-done-ili-vse-chto-sleduet-znat-o-zhelaniyah-polzovateley (last accessed 2025/08/29). (In Russian)
28. Prakasha, G.S., Kenneth, A., Lapina, M. Social media disorder in relation to device screen time and autosuggested social media content engagement. Závislost způsobená časem stráveným na sociálních sítích v souvislosti s automaticky generovaným obsahem na sociálních sítích Kontaktthis, 2023, 25(2), pp. 117–123, https://doi.org/10.32725/kont.2023.021.

Comparative Analysis of Deep Learning and Ensemble Methods for Intrusion Detection in IoT Networks

Melsa Magdalene Fernandes, Sharlene Anna Pereira, Mary Anita E A(✉), and Vinodha D

Christ University, Bangalore, India
maryanita.ea@christuniversity.in

Abstract. The increasing number of Internet of Things (IoT) devices has tremendously increased the vulnerability of today's networks to a wide range of sophisticated cyber-attacks. Traditional legacy intrusion detection systems (IDS) that typically employ rule-based or signature-based systems do not tend to identify new or emerging attacks because they are static systems. Considering these limitations, this paper suggests an ensemble-based intrusion detection system through deep learning and machine learning for better adaptability and resiliency. The paper explains a range of deep learning architectures such as Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Artificial Neural Networks (ANN), and Long Short-Term Memory (LSTM) networks that have the capability of learning spatial, temporal, and complex patterns of network traffic data. Complementing these strengths, CNN is utilized for feature representation purposes, whereas Random Forest (RF) and XGBoost are utilized at the decision level to build a heterogeneous ensemble. This hybrid model exploits the strength of CNN to learn spatial correlations as well as interpretability and robustness of tree classifiers. The suggested framework offers a balanced and responsive IDS that can solve problems like high-dimensional features, class imbalance, and minority attack category detection.

Keywords: Hybrid IDS · Deep Learning IDS · CNN RF XGBoost IDS · Ensemble Intrusion Detection · Feature Selection IDS · SMOTE IDS · Class Imbalance IDS · Adaptive Intrusion Detection · Semi-supervised IDS · Unsupervised IDS · Adversarial Attack Detection · Evasion Attack Resistance · Cybersecurity IDS · Hybrid Deep Learning Security

1 Introduction

IoT device proliferation greatly increases the functionality of today's networks while simultaneously making them vulnerable to an increased list of cyber-attacks. Due to their heterogeneity, minimal processing capacities, and use of varying communication protocols, IoT networks are particularly vulnerable to Distributed Denial of Service (DDoS) attacks, reconnaissance, and data loss. Rule or signature based conventional

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 103–111, 2027.
https://doi.org/10.1007/978-3-032-27834-0_10

intrusion detection systems (IDS) are often not able to detect new or evolving threats, thereby highlighting the need for adaptive and smart IDS solutions. The recent advances in machine learning (ML) and deep learning (DL) have promised wonderful solutions to these problems. Architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks can learn automatically spatial and temporal patterns from traffic data, while traditional ML models like Random Forest (RF) and XGBoost ensure stable and explainable decision making. Nonetheless, their performance is impaired by typical network data set problems such as high dimensionality, noisy attributes, and extreme class imbalance with disproportionate impact on minority but highly significant attack types. Furthermore, no model consistently produces strong performance on all types of traffic.

To counter these limitations, this research suggests a heterogeneous ensemble-based IDS framework that combines CNN for feature learning with RF and XGBoost as decision-level classifiers. This heterogeneous framework exploits the representation learning capability of deep neural models with the resilience of ensemble tree-based methods, providing balanced detection performance between majority and minority classes. The contributions of this work are two-fold: (i) a comparative evaluation of popular DL models for IDS applications, and (ii) the proposal of a scalable hybrid ensemble approach that improves adaptability and reliability for heterogeneous IoT environments.

Cyber-attacks pose a huge threat, as they can lead not only to financial losses, but also to reputational damage and legal consequences [2, 6, 11, 12]. In this regard, information security issues are extremely relevant today.

2 Materials and Methods

2.1 Experimental Workflow

The IDS is developed sequentially step by step, beginning with training and developing various deep learning models like CNN, DNN, ANN, and LSTM on the Bot-IoT and UNSW-NB15 datasets for performance benchmarking. This provides a comparative study of the feature extraction capability of each model and network traffic classification capability. Based on this comparison, CNN is selected as the optimal deep learning model for feature extraction since it possesses the advanced capability to extract spatial correlation among network traffic features. Subsequently, different ensemble methods are attempted by combining CNN with other learners such as LSTM, DNN, Logistic Regression, SVM, Random Forest (RF), and XGBoost. Ensemble models are then trained on the Bot-IoT dataset to verify it's fit for large scale testing. These experiments highlight the complementary strengths of individual learners and guide the proposed hybrid design.

2.2 Datasets and Preprocessing

Two benchmark datasets are employed in this work: Bot-IoT, having IoT specific traffic with both benign and attack flows, and UNSW-NB15, having a wide range of attack types. The raw data is then preprocessed for the model.

- **Data Cleaning** Missing values in columns are eliminated, and small inconsistencies are normalized.

- **Categorical Encoding** The target variable and non-numeric features are encoded into numbers to make them model compatible.
- **Class Balancing** Against class imbalance, Random Undersampling reduces majority class dominance and is complemented with SMOTE (Synthetic Minority Oversampling Technique) for oversampling minority classes. Fig. 1 shows the dataset distribution prior to balancing as well as after balancing, where the enhanced proportionality of attack classes is emphasized.

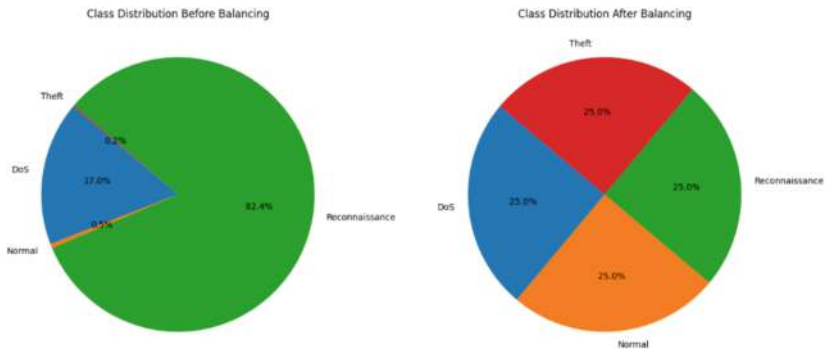


Fig. 1. Class Distribution before and after balancing

- **Feature Scaling and Splitting** RobustScaler scales the features and the data is split into train and test sets with 80% and 20%, respectively, via stratified sampling to maintain class distribution.

2.3 Proposed Hybrid Model

The end system combines the power of deep learning and ensemble methods, CNN + RF + XGBoost.

- CNN Feature Extraction The preprocessed data is trained with a 1D Convolutional Neural Network so that it can learn network feature spatial relationships. Activation of this dense feature layer is utilized as learned embeddings.
- RF and XGBoost Classification The CNN embeddings are input to Random Forest and XGBoost classifiers. RF provides class imbalance robustness, and XGBoost learns complex non-linear relationships in the data.
- Ensemble Decision Layer The final prediction probabilities for each class are then obtained by averaging the CNN, RF, and XGBoost outputs using soft voting and thereby improves the overall stability and accuracy of detection.

3 Results

Bot-IoT and UNSW-NB15 datasets are utilized to experiment with the proposed framework. To assess detection abilities on different classes of attacks, individual deep learning models as well as ensemble settings are tested. Apart from classwise analysis

to emphasize detection performance for rare but important attacks such as theft and reconnaissance, the evaluation metrics include accuracy, precision, recall, and F1-score.

3.1 Performance of Individual Models

When taken individually, the performance of each deep learning approach shows that CNN outperforms the other approaches on both datasets. With ~98.8% accuracy on Bot-IoT and 96% accuracy on UNSW-NB15, CNN is effective in capturing spatial relationships between traffic features, which is beneficial for structured network data. DNN also performs similarly in terms of accuracy, but its recall and F1-scores are significantly lower, implying that it classifies majority classes well but performs poorly on minority classes. LSTM models, which are intended to model sequential dependencies, perform fairly well on Bot-IoT but exhibit much weaker performance on UNSW-NB15, showing sensitivity to the dataset and poor performance of temporal modeling when minority types of attacks are poorly represented (Table 1).

Table 1. Performance of individual models across Bot-IoT and UNSW datasets.

Model	Dataset	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)
CNN	Bot-IoT	~98.8%	0.73	0.52	0.53
DNN	Bot-IoT	~99.8%	0.57	0.50	0.50
ANN	Bot-IoT	~82%	0.08	0.10	0.09
LSTM	Bot-IoT	~91%	0.66	0.40	~0.90 (weighted)
CNN	UNSW	96%	0.99	0.97	0.98
DNN	UNSW	~89%	0.66	0.40	0.32
ANN	UNSW	~80%	0.08	0.10	0.09
LSTM	UNSW	~79.8%	0.21	0.25	0.23

3.2 Performance of Ensemble Models

CNN + DNN has the best raw accuracy at 97.95%, but its improved test loss and sporadic minority class misclassifications indicate that its predictions are not as consistent. CNN with simple classifiers such as Logistic Regression and SVM also works great, both with high recall values. CNN + Logistic Regression particularly stands out with its simplicity of approach and equal detection across classes, whereas CNN + SVM also shows resilience with slightly lower recall for rare attacks.

The suggested CNN + RF + XGBoost ensemble model results in an accuracy of 97.32% with the minimum test loss (0.0658). Most notably, it demonstrates ideal recall for minority attack classes, making it extremely efficient in identifying rare but potent attacks like Theft. Through the synergistic combination of CNN's capability to learn deep features and the robustness of tree-based learners, this ensemble obtains a balanced trade-off between accuracy, stability, and generalizability. While its accuracy is marginally less

than the CNN + DNN hybrid model, its greater robustness and minority class detection feature render it the most realistic and feasible option for practical intrusion detection deployments (Table 2).

3.3 Visualization of Proposed Model Performance

Table 2. Performance of ensemble models.

Model	Accuracy	Test loss	Minority class recall (theft)	Key observations
CNN + LSTM	91.42%	0.5963	0.71	Captures temporal features, limited recall on rare classes
CNN + DNN	97.95%	0.1447	0.94	Highest accuracy, but unstable for minority classes
CNN + Logistic Regression	97.17%	0.1132	1.00	Balanced across classes, lightweight but less flexible
CNN + SVM	96.77%	0.5170	0.99	Good generalization, but poorer recall
CNN + RF + XGBoost	97.32%	0.0658	1.00	Lowest loss, very stable, ideal recall on rare attacks

To further validate the proposed CNN + RF + XGBoost ensemble, some visualization is explored along with the quantitative analysis. These visual results represent compelling evidence for the robustness of the model and identify its capability to generalize on both minority and majority classes. The classification report (Fig. 2) clearly illustrates the well-balanced performance of the ensemble on all four classes. DoS and Normal traffic are accurately labeled with almost perfect precision and recall (0.99–1.00), and Reconnaissance is labeled with a very good F1-score of 0.95 due to its poor recall (0.90). In fact, the infrequent Theft category, usually the most difficult to label, is also accurately labeled with perfect recall (1.00) and an F1-score of 0.95. The weighted average and macro also validate the fact that the model is overall precise to 0.98, recall to 0.97, and F1-score to 0.97, portraying superb generalization with less bias towards majority classes.

Ensemble Accuracy: 0.9732, Loss: 0.0658

Classification Report:

	precision	recall	f1-score	support
DoS	0.99	0.99	0.99	318
Normal	1.00	1.00	1.00	317
Reconnaissance	1.00	0.90	0.95	318
Theft	0.91	1.00	0.95	317
accuracy			0.97	1270
macro avg	0.98	0.97	0.97	1270
weighted avg	0.98	0.97	0.97	1270

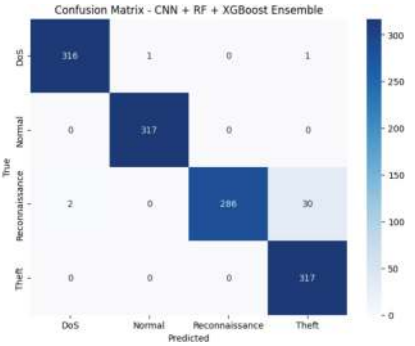


Fig. 2. Classification report and **b** confusion matrix.

The confusion matrix gives a clearer look at the performance of classification. A very small number of samples (a handful from the total 1,270) are mislabeled. For instance, two Reconnaissance examples are labeled as DoS and 30 as Theft, and one example of DoS is wrongly marked as Normal. Despite these small imperfections, diagonal dominance of the matrix confirms that the ensemble is able to distinguish between different types of attacks and normal traffic consistently. Most importantly, Theft cases are misclassified with zero misclassifications, further proving the efficiency of the ensemble in identifying key but poorly represented attack classes.

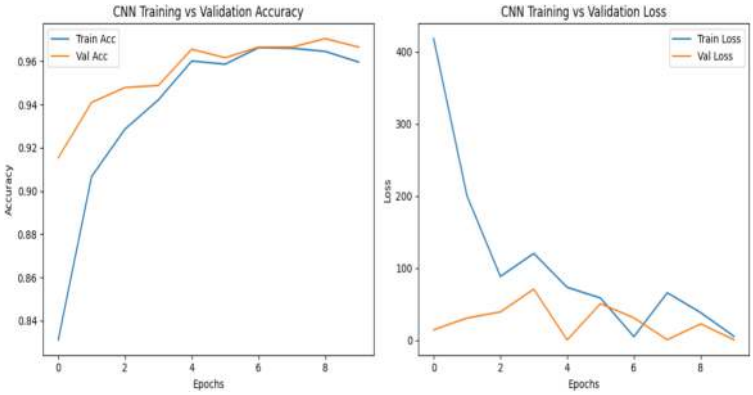


Fig. 3. Accuracy and loss graphs.

Training and validation curves (Fig. 3) give us a sense of the learning pattern of the CNN in the ensemble. The graph of accuracy presents a consistent rise for training as well as validation sets, with validation accuracy marginally higher than training accuracy and plateauing at over 96% from the sixth epoch onwards. Similarly, the training loss plot also depicts a sharp drop in training loss across epochs, while validation loss remains always low and showing the same trend. The reason that there is not significant divergence of the two curves indicates that the model avoided overfitting and led to high accuracy as well as strong generalization on new data. Collectively, these visualizations support

the quantitative results by confirming that the suggested ensemble of CNN + RF + XGBoost is very accurate and robust.

4 Discussion

The success of the CNN + RF + XGBoost model highlights the power of hybrid architectures, an emergent topic among recent IDS research. Overall overviews of the subject matter, as presented by Saranya et al. [9], uniformly name tree-based models such as Random Forest among the highest performing classifiers and thus warrant its presence in our ensemble. The theory that it is better to combine models than use separate models is illustrated by Sridevi et al. [10], whose voting classifier outperforms its constituents, and by Fu [4] and Rani et al. [8], who both achieve accuracies greater than 98% by combining deep neural networks with Random Forest. Our application of a CNN to this task is supported by Musleh et al. [7], finding that a VGG-16 deep learning feature extractor is crucial in order to achieve high accuracy in an IoT scenario. The need to tackle class imbalance using methods such as SMOTE and the use of resilient feature selection is emphasized by Bakro et al. [1] in their cloud-based IDS. The influence of feature selection on classifiers such as RF and XGBoost is also explored further by Zouhri et Al [14].

Although the performance of the proposed model is very competitive, it is essential to place these results in the general context of operational challenges and competing methodologies. Detection of advanced threats is a war that needs to be fought continuously, and as Kim and Pak [5] show, well trained machine learning models themselves can be susceptible to adversarial attacks that introduce feature values beyond the scope of training data. Their contribution on "early classification" points towards the requirement of dynamic and resilient defense strategies over static training. In addition, our supervised method depends on having attack data that is properly labeled, a large practical limitation. This problem motivates the investigation of alternative learning paradigms. For example, Wang et al. [13] build a semi supervised system employing OCSVM and GMM requiring only normal data for training. Likewise, Bhadauria and Mohanty [3] propose a hybrid system that includes an unsupervised module purely to identify new attacks without any signatures. These works indicate that although our model is extremely accurate for its intended purpose, the way ahead in intrusion detection lies in developing adaptive systems capable of handling new evasion strategies and working effectively under conditions with limited labeled data.

5 Conclusion

This work proposes a hybrid ensemble intrusion detection system that enhances CNN-based feature extraction with Random Forest and XGBoost based classification. By integrating deep learning with tree-based learners, the approach balances detection for various attack categories, including minority threats, which usually get disregarded by individual models. Benchmark dataset evaluation demonstrates its scalability and resilience in heterogeneous IoT scenarios, rendering it an efficient remedy against the weaknesses of conventional and single model IDS solutions. The proposed framework

offers a foundation for developing adaptive IDS strategies and can be extended to support real-time security in emerging IoT applications.

References

1. Bakro, M., et al.: An improved design for a cloud intrusion detection system using hybrid features selection approach with ML classifier. IEEE Access. **11**, 64228–64247 (2023). https://doi.org/10.1109/ACCESS.2023.3289405
2. Basan, E., Makarevich, O., Lapina, M., Mecella, M.: Analysis of the Impact of a GPS Spoofing Attack on a UAV CEUR Workshop Proceedings. **3094**, 6–16 (2022) https://ceur-ws.org/Vol-3094/invited_paper.pdf
3. Bhadauria, S., Mohanty, T.: Hybrid intrusion detection system using an unsupervised method for anomaly-based detection. In: Proceeding 2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1–6, Hyderabad, India (2021). https://doi.org/10.1109/ANTS52808.2021.9936919
4. Fu, R.: Design and implementation of network intrusion detection system based on machine learning. In: Proceeding 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN), pp. 1–6, Bidar, India (2025). https://doi.org/10.1109/ICISCN64258.2025.10934502
5. Kim, T., Pak, W.: Robust network intrusion detection system based on machine learning with early classification. IEEE Access. **10**, 10754–10767 (2022). https://doi.org/10.1109/ACCESS.2022.3145002
6. Lapina, M., Sajid, M., Bagautdinova, A., Lapin, V., Tokmakova, M. (2024). Comparative Characteristics of the Methods of Using Digital Watermarks to Protect Text Information. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds) AISMA-2023: International Workshop on Advanced Information Security Management and Applications. AISMA 2023. Lecture Notes in Networks and Systems, vol 1207. Springer, Cham. doi:https://doi.org/10.1007/978-3-031-77229-0_15
7. Musleh, D., Alotaibi, M., Alhaidari, F., Rahman, A., Mohammad, R.M.: Intrusion detection system using feature extraction with machine learning algorithms in IoT. J. Sens. Actuator Netw. **12**(2), 29 (2023). https://doi.org/10.3390/jsan12020029
8. Rani, B.P., Lahari, C.S., Lavanya, J.A., Lakshmanarao, A., Kosuri, G.V.: Advanced intrusion detection system through hybrid integration of Random Forest and deep learning with artificial neural networks. In: Proceeding 2024 3rd International Conference for Advancement in Technology (ICONAT), pp. 1–5, Goa, India (2024). https://doi.org/10.1109/ICONAT61936.2024.10774823
9. Saranya, T., Sridevi, S., Deisy, C., Chung, T.D., Khan, M.K.A.A.: Performance analysis of machine learning algorithms in intrusion detection system: a review. Procedia Comput. Sci. **171**, 1251–1260 (2020). https://doi.org/10.1016/j.procs.2020.04.133
10. Sridevi, S., Prabha, R., Reddy, K.N., Monica, K.M., Senthil, G.A., Razmah, M.: Network intrusion detection system using supervised learning-based voting classifier. In: Proceeding 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT) (2022). https://doi.org/10.1109/ic3iot53935.2022.9767903
11. Tchernykh, A. et al.: Towards Mitigating Uncertainty of Data Security Breaches and Collusion in Cloud Computing, pp. 137–141 (2017). https://doi.org/10.1109/DEXA.2017.44
12. Vinodha, D., Buvana, M.J., Rajalakshmi, S., Anita, E.A.M., Lapina, M.: Framework based on IoT, AI, and blockchain for smart access to government agricultural schemes, The Future of Agriculture: IoT. AI Blockchain Technol. Sustain. Farming, 142–163 (2024). https://doi.org/10.2174/9789815274349124010011

13. Wang, C., Sun, Y., Lv, S., Wang, C., Liu, H., Wang, B.: Intrusion detection system based on one-class support vector machine and Gaussian mixture model. Electronics. **12**(4), 930 (2023). https://doi.org/10.3390/electronics12040930
14. Zouhri, H., Idri, A., Ratnani, A.: Evaluating the impact of filter-based feature selection in intrusion detection systems. Int. J. Inf. Secur. **23**, 759–785 (2024). https://doi.org/10.1007/s10207-023-00767-y

Actualization of the "Hero Image" as a Consequence of the Convergence of Artificial Intelligence and Telecommunication Systems

Oleg Gabrielyan[1], Yelizaveta Vitulyova[2,3,4(✉)], and Ibragim Suleimenov[3]

[1] V.I. Vernadsky Crimean Federal University, Simferopol, Crimea 295007, Ukraine
[2] National Scientific Laboratory for the Collective Use of Information and Space Technologies (NSLC IST), Almaty 050013, Kazakhstan
lizavita@list.ru
[3] Engineering Technological University, Almaty 050060, Kazakhstan
[4] JSC Institute of Digital Engineering and Technology, Almaty 050013, Kazakhstan

Abstract. It is demonstrated that the pronounced development of telecommunication networks significantly increases the role of the collective component of human consciousness, which has a dual nature. Qualitative transformations of human consciousness are predicted, including those resulting from the convergence of artificial intelligence and telecommunication networks. It is shown that the trajectory of human consciousness evolution, as well as the direction of artificial intelligence development, is currently determined by factors that are random in nature. At the same time, there exists a wide range of possible evolutionary scenarios, including those of a clearly negative character. To prevent such outcomes, it is advisable to develop tools that enable forecasting. One such tool involves identifying hero images of various types, including at the early stages of the emergence of corresponding informational objects evolving at a supra-individual level of information processing. This requires a consistent, scientifically grounded interpretation of the term "hero image," which is provided on the basis of an analogy with ancient myths. It is shown that such myths can serve as sources of insight into the evolution of human consciousness at its early stages, since ancient deities can be viewed as distorted reflections of objects evolving at the supra-individual level of information processing.

Keywords: Artificial Intelligence · Telecommunication Systems · Consciousness Duality · Hero Image · Noosphere · Supra-Individual Information Processing

1 First Section

In our previous works [1, 2], we have demonstrated that human consciousness possesses a dual nature. This idea can already be traced in the non-classical psychology of L. S. Vygotsky [3, 4], but the neural network model of the noosphere, initially proposed in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 112–122, 2027.
https://doi.org/10.1007/978-3-032-27834-0_11

[5, 6], allows for a consistent natural-scientific justification of such views. The duality of consciousness implies the simultaneous presence of both individual and collective components, which can be demonstrated even without recourse to mathematical models, one of which was proposed in our study [7].

Consciousness as a physical object does not, in fact, exist; physically, only the exchange of electrochemical signals between neurons of the brain exists. Consciousness is the product of processing these signals. Similarly, interpersonal communication is also reduced to the exchange of signals between neurons, but belonging to different individuals.

The difference between intracerebral signal exchange and interpersonal communication lies in the fact that in the former case neurons are directly connected to each other, while in the latter the exchange of signals occurs sporadically through additional communication channels: acoustic, optical, tactile. However, as follows from the basic principles of neural network theory [8, 9], the functioning of a network as a whole does not require the communication channels between neurons to exist in a stationary manner. It is sufficient that they arise at the moment of signal transmission, which is exactly what occurs in interpersonal communication.

As a result, human society in fact forms a global neural network, which, with certain reservations, can be identified with the noosphere as understood by V. I. Vernadsky [10, 11]. Consequently, the individual brain—or more precisely, the fragment of the global neural network it forms—constitutes only a part of this larger system. Figure 1 illustrates the schematic representation of the formation of a common neural network resulting from interpersonal communication.

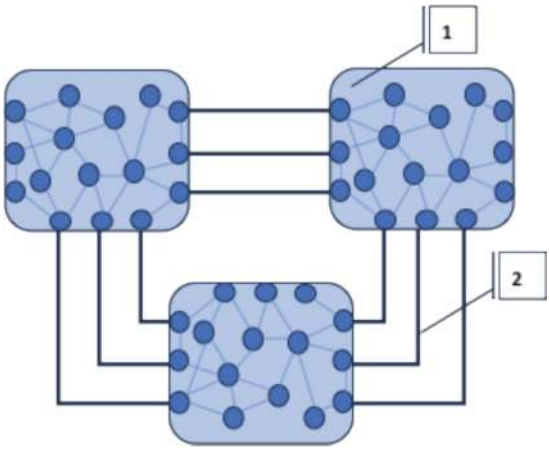


Fig. 1. Illustration of the formation of a common neural network as a result of interpersonal communication. 1—schematic representation of the neural network formed by an individual’s brain; 2 – schematic representation of communication channels between individuals (acoustic, optical, etc.)

2 Problem Statement

This very circumstance leads us to assert that human consciousness possesses a dual nature [1, 2]. Since signal exchange within the brain is significantly more intense than that associated with interpersonal communication, human consciousness is indeed relatively autonomous, giving rise to the individual component.

At the same time, this fragment of the neural network remains part of the global network, which leads to the formation of the collective component of consciousness [12]. Clearly, with the development of telecommunication technologies, the intensity of information exchange between individuals continues to increase. Thus, it is safe to say that society has long since been transformed de facto into a human–machine system.

Moreover, neural network theory unequivocally indicates that the informational capacity of a network depends nonlinearly on the number of its elements. Otherwise, there would be no sense in creating increasingly large neural networks to solve practical tasks [13, 14]. This means that the global neural network is capable of generating qualitatively new phenomena – those not inherent to individual human beings, even in aggregate.

We emphasize that a new quality arises in the philosophical sense of the term. We interpret it as a supra-individual level of information processing, at which objects of various natures may emerge and evolve (Fig. 2).

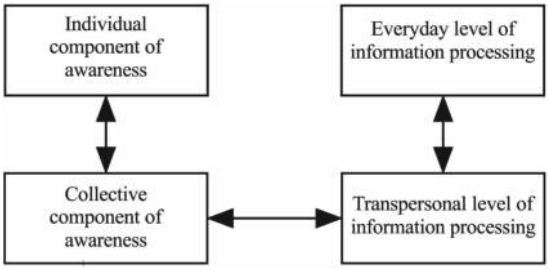


Fig. 2. On the formation of the supra-individual level of information processing.

Moreover, in our view, the evolution of consciousness represents the co-evolution of its individual and collective components, with one or the other dominating at different historical periods [12] (see Fig. 3).

The general theory of the evolution of complex systems, particularly as presented in our work [15], allows us to assert that during the formative stage of consciousness, the collective component prevailed. The individual component became dominant significantly later, presumably during the period marking the end of matriarchy.

The transition from one dominant form to another inevitably led to an intermediate state in which both components of consciousness were relatively balanced (see Fig. 3). This suggests that individual consciousness could reflect processes occurring at the supra-individual level of information processing. This very conclusion provides the basis for a justified hypothesis: ancient deities represented informational objects evolving at the supra-individual level of information processing.

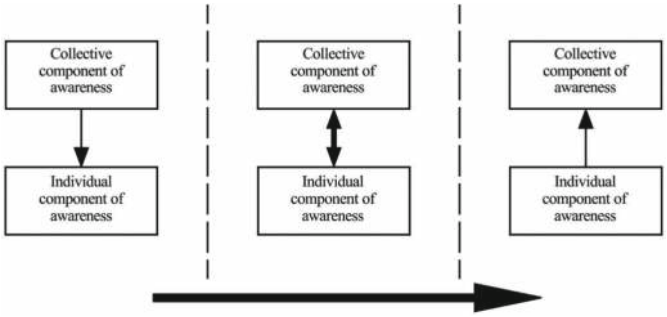


Fig. 3. On the correlation between the significance of the individual and collective components of consciousness across different historical periods.

Individual consciousness of that era was incapable of systematically analyzing such information, and thus the resulting image was inevitably distorted. Supra-individual informational structures or objects were perceived by our distant ancestors in the form of ancient deities.

3 Conceptual Framework

Given that society has been converted into a human–machine system and interpersonal information exchange is becoming increasingly intensive, there are strong grounds to believe that the influence of the collective component of consciousness will continue to grow. Consequently, in the foreseeable future, questions surrounding various forms of interaction with informational objects evolving at the supra-individual level of information processing are likely to arise.

Furthermore, in the near future, it is reasonable to expect that entities developing at the supra-individual level of information processing will exert increasing influence on society. This circumstance compels us to consider ancient myths as an important – albeit significantly distorted – source of information. At the very least, one can argue that myths metaphorically reflect real processes associated with the evolution of consciousness.

This conclusion is supported by a widely recognized fact: structurally, the myths of various cultural peoples – even those who never came into contact with one another – are strikingly similar [16]. Such a reading of myths draws special attention to the figure of the hero, who appears in the narratives of the vast majority of cultures worldwide. In myths, the hero acts as a kind of mediator between the supra-individual level of information processing and the ordinary one that is familiar to all of us.

According to mythological sources, the distinctive feature of the hero is the ability to interact directly with the gods. For example, in ancient Greek myths, many heroes are said to be direct descendants of the gods. In other words, the hero, as a mediator, is the heir, interpreter, or herald of certain supra-individual informational structures, which our distant ancestors interpreted as ancient gods. Fig. 4 presents the role of the hero as a mediator between the supra-individual and individual levels of information processing.

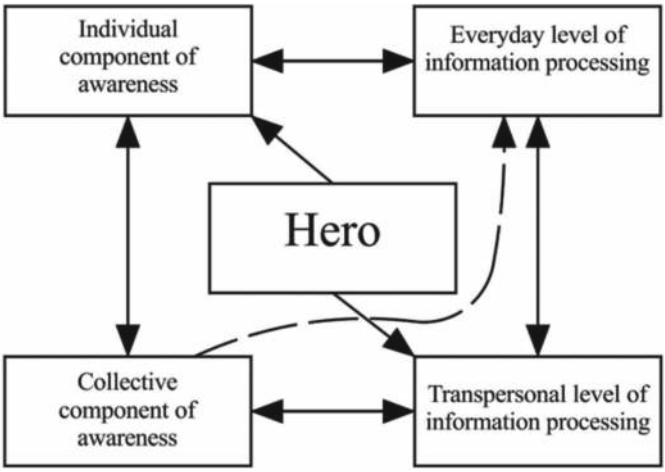


Fig. 4. The Hero as a Mediator between Supra-Individual and Individual Levels of Information Processing.

It is evident that with the development of telecommunication systems and the strengthening of the collective component of consciousness, one can reasonably anticipate the emergence of new mediators who will fulfil the same function as the heroes of classical ancient Greek mythology – albeit with appropriate modifications. In a certain sense, history may repeat itself, but on a new turn of the spiral. This prompts us to reflect on who or what may become the hero of modern society in the sense outlined above.

4 Methodological Considerations

Who is capable of assuming the role of a mediator between the ordinary and the supra-individual levels of information processing? In this context, it is particularly important to note that modern telecommunication systems are effectively converging with artificial intelligence systems. In other words, AI can already be considered a means – albeit a primitive one – of establishing contact with the supra-individual level (see Fig. 5).

Indeed, it is extremely difficult for any of us to comprehend the processes occurring at the supra-individual level of information processing – just as it is difficult to imagine a dialogue between a person and a single neuron in their own cerebral cortex. These are fundamentally different levels of organization in information processing systems.

However, if a tool capable of analyzing processes across the entire telecommunication environment emerges, then the possibility of establishing a certain form of contact with entities arising at this level becomes real. Of course, direct data collection alone will yield little. A systemic theory is required to interpret the resulting experimental data. This is precisely what is illustrated in Fig. 5, which largely mirrors Figure 4. The difference lies in the fact that, at present, AI cannot serve as a full-fledged mediator – an interpreter is required, which, unfortunately, may not necessarily be the scientific or technical community. The results of AI usage may be co-opted by a wide range of actors,

including counter-elite and anti-elite groups, for whom understanding the essence of the processes may become secondary.

Figure 5 illustrates the use of artificial intelligence as a mediator between the supra-individual and ordinary levels of information processing at the current stage.

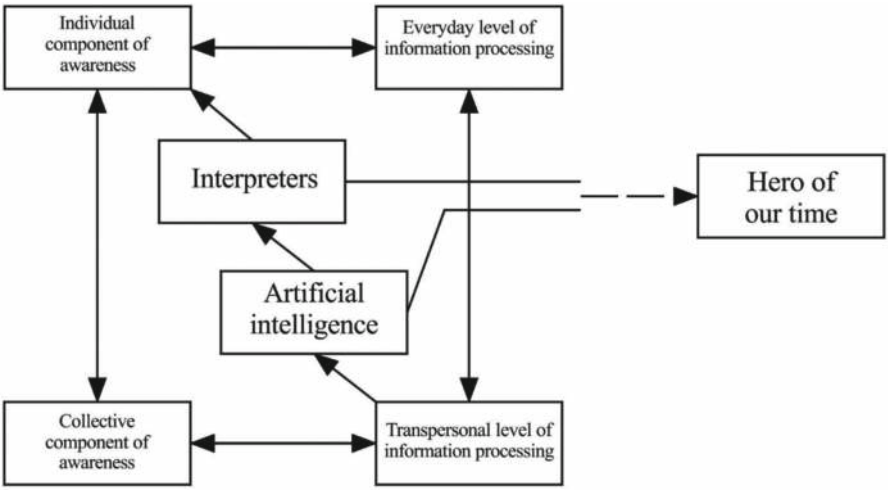


Fig. 5. On the use of AI as a mediator between the supra-individual and ordinary levels of information processing at the current stage.

5 Discussion

Nevertheless, the potential for such contact may arise through natural processes of self-organization within the telecommunication system that has converged with AI. In this regard, it is worth noting that the literature already includes works discussing the possibility of spontaneous intelligence emerging within the global telecommunication network [17, 18], as well as advancements in hierarchical intelligent systems that could facilitate such processes [19]. Naturally, this raises numerous complex questions, primarily what should be understood by the term "intelligence".

As noted in [20], most definitions of intelligence in existing literature are descriptive in nature. They characterize the only reliably known form of intelligence – the human one. However, even at this stage of research, it can be argued that intelligence should primarily be understood as a system of information processing [20], and in this capacity, spontaneous intelligence can undoubtedly arise.

Moreover, entities of a much more complex nature may emerge, which can be interpreted as heroes in the sense described above. It is important to emphasize that a hero, strictly speaking, does not have to be a specific human being. It may be a generalized or collective image – more precisely, an informational object originating at the supra-individual level of information processing but still accessible to the perception of ordinary individuals.

Such an object must, of course, be personified in some way – as occurred in antiquity. For this reason, it can be regarded as a personified hero, embodied in the form of a collective image (e.g., literary or otherwise).

There is reason to believe that certain prototypes of heroes in the sense described – although not yet fully formed – already exist within the information space. With some caveats, they can be called "embryonic" heroes, by analogy with ancient myths in which heroes served as mediators between the individual and supra-individual levels of information processing.

Identifying such "embryonic" heroes appears to be of critical importance, as it may enable forecasting related to the transformation of human consciousness. The conclusion that the noosphere has already been converted into a human–machine system, along with the strengthening of the collective component of consciousness, allows us to assume that in the foreseeable future humanity may face changes far more profound than those brought about by the industrial revolutions or shifts in historical phases. While past transformations did not affect the nature of consciousness itself, the coming changes may lead to a qualitatively new state of consciousness that currently defies accurate prediction.

Furthermore, a wide range of scenarios may unfold – from highly positive to extremely adverse. The positive scenario entails the enhancement of human cognitive abilities; the negative one suggests dehumanization, the dissolution of individual consciousness into the collective mass, and the actual subjugation of the individual by supra-individual informational entities, including artificially created ones. In this context, it is worth emphasizing that the scenarios for the evolution of consciousness clearly correlate with the scenarios for the development of artificial intelligence, which are themselves highly variable [21].

This domain also exhibits both positive and negative trajectories with comparable consequences. It is evident that modern civilization is at a bifurcation point: the paths for AI development remain undefined, under discussion, and, for the most part, are unfolding spontaneously. AI is being developed by major corporations, political elites, and other actors in response to tactical objectives, without a systemic strategic vision.

Moreover, current discussions remain fragmented and have not yet evolved into a general scientific discourse. If this situation persists, the trajectory of AI development – and, consequently, the direction of human consciousness evolution—will be determined randomly, through a mechanism of spontaneous selection.

As the theory of scientific revolutions shows [22, 23], when multiple competing directions exist in the development of a technical system, one dominant vector eventually prevails. A typical example is the history of computing: in the 1970s, alongside binary logic machines, ternary logic machines also existed [24].

Despite the obvious advantages of ternary logic—including higher information density and potential efficiency, as reviewed in [25]—binary logic became the dominant approach, largely due to economic and infrastructural factors. Today, ternary logic is almost entirely absent from computing, although a shift to a different hardware base could theoretically revive its relevance [26]. This example illustrates the central thesis: the mainstream typically emerges along only one of several possible paths.

This is precisely why it seems more than necessary to ensure at least the development of tools capable of monitoring the current state of affairs. In this context, the creation of mechanisms capable of identifying what has previously been termed “embryonic heroes” becomes particularly important. Such tools can clearly be based on artificial intelligence systems that monitor the flow of information circulating within telecommunication networks, including machine learning techniques for managing network quality of service [27].

In a certain sense, this type of tool also functions as a mediator between the ordinary and supra-individual levels of information processing. The underlying operational principles of such systems may vary widely. However, discussion of these principles falls outside the scope of this study.

Our objective is different. We aim to demonstrate the significance of this problem, while its resolution may be approached through a variety of methods, including those based on the artificial synthesis of neural networks using operations over Galois fields [28] or finite algebraic rings [29].

6 Conclusion

Metaphorically, our perspective is reflected in Fig. 6. The informational entities (possibly personified) that ancient mythology interpreted as heroes may return in a new form, and it is crucial to identify their emergence at the earliest stages of their development.

Fig. 6 metaphorically represents the role of the hero as a mediator between the supra-individual and ordinary levels of information processing.

Thus, the hero—based on the interpretation of ancient mythology proposed here—should be understood as a kind of mediator between the individual and supra-individual levels of information processing. The emergence of such heroes (not necessarily personified) may be seen as an indicator of the increasing influence of the collective component of consciousness. Consequently, identifying the emergence of “heroes” at various stages of this process can serve as a tool for analyzing the dynamics occurring at the supra-individual level of information processing and, by extension, for developing informed forecasts.

The development of such tools is highly relevant today, as civilization currently stands at a bifurcation point – a fact that underscores the need for a systematic analysis of the implications arising from the widespread deployment of artificial intelligence.

Fig. 6. Metaphorical representation of the hero as a mediator between the supra-individual and ordinary levels of information processing.

Acknowledgments. The study was carried out with the support of the Russian Science Foundation grant (project No. 24-28-00413 "The Image of the Hero in the Contemporary Russian Media Space: Current State and Prospects").

References

1. Vitulyova, Y., Gabrielyan, O., Bakirov, A., Suleimenov, I.: Humanist ideals in an era of increasing confrontation: the need to renew basic paradigms. J. Ecohumanism. **3**(7), 2064–2076 (2024)
2. Suleimenov, I., Gabrielyan, O., Matrassulova, D.: Philosophical foundations of sciences and prospects of multivalued logic in describing thinking. Sci. & Educ. **1–19** (2025)
3. Robbins, D.: Vygotsky's non-classical dialectical metapsychology. J. Theory Soc. Behav. **33**(3), 303–312 (2003)
4. Hakkarainen, P.: Towards nonclassical methodology of psychology. J. Russ. East Eur. Psychol. **52**(5–6), 1–12 (2015)
5. Suleimenov, I.E., Gabrielyan, O.A., Massalimova, A.R., Vitulyova, Y.S.: World spirit from the standpoint of modern information theory. Eur. J. Sci. Theol. **20**(1), 19–31 (2024)
6. Suleimenov, I., Massalimova, A., Bakirov, A., Gabrielyan, O.: Neural networks and the philosophy of dialectical positivism. MATEC Web Conf. **214**, 02002 (2018)
7. Suleimenov, I.E., Matrassulova, D.K., Moldakhan, I., Vitulyova, Y.S., Kabdushev, S.B., Bakirov, A.S.: Distributed memory of neural networks and the problem of the intelligences essence. Bull. Electr. Eng. Inform. **11**(1), 510–520 (2022)
8. Goodfellow, I. et al.: Generative adversarial networks. Commun. ACM. **63**(11), 139–144 (2020). https://doi.org/10.1145/3422622

9. Fine, T.L.: Fundamentals of artificial neural networks [book reviews]. IEEE Trans. Inf. Theory. **42**(4), 1322 (1996). https://doi.org/10.1109/TIT.1996.508868
10. Jasečková, G., Konvit, M., Vartiak, L.: Vernadsky’s concept of the noosphere and its reflection in ethical and moral values of society. Hist. Sci. Technol. **12**(2), 231–248 (2022)
11. Oldfield, J.D., Shaw, D.J.: Vernadsky and the noosphere concept: Russian understandings of society–nature interaction. Geoforum. **37**(1), 145–154 (2006)
12. Ibragim, S., Oleg, G., Vitulyova, Y.: Problems of many-valued logic from the point of view of the theory of socio-cultural code. J. Ecohumanism. **3**(4), 236–248 (2024)
13. Kaplan, J. et al.: Scaling laws for neural language models. https://arxiv.org/abs/2001.08361 [cs.LG] (2020). https://doi.org/10.48550/arXiv.2001.08361
14. Hestness, J. et al.: Deep learning scaling is predictable, empirically. https://arxiv.org/abs/1712.00409 [cs.LG] (2017). https://doi.org/10.48550/arXiv.1712.00409
15. Kabdushev, S., Gabrielyan, O., Kopishev, E., Suleimenov, I.: Neural network properties of hydrophilic polymers as a key for development of the general theory of evolution. R. Soc. Open Sci. **12**(4), 242149 (2025)
16. Ben-Amos, D.: Introduction to the special issue “the challenge of folklore to the humanities”. Humanities. **10**, 18 (2021). https://doi.org/10.3390/h10010018
17. Elek, I., Roden, J., Nguyen, T.B.: Spontaneous emergence of the intelligence in an artificial world. In: International Conference Neural Information Processing, pp. 703–712 (2012)
18. Chen, J., Burgess, P.: The boundaries of legal personhood: how spontaneous intelligence can problematise differences between humans, artificial intelligence, companies and animals. Artif. Intell. Law. **27**(1), 73–92 (2019)
19. Velyamov, T., Kim, A., Manankova, O.: Modification of the Danzig-wolf decomposition method for building hierarchical intelligent systems. Int. J. Adv. Comput. Sci. Appl. **15**(7), 1160–1167 (2024)
20. Suleimenov, I.E., Vitulyova, Y.S., Bakirov, A.S.: Gabrielyan, O.A.: artificial intelligence: what is it? In: Proceedin 6th International Conference, pp. 22–25. Computer and Technology Applications (2020)
21. Kalimoldayev, M.N., Pak, I.T., Baipakbayeva, S.T., Mun, G.A., Shaltykova, D.B., Suleimenov, I.E.: Methodological basis for the development strategy of artificial intelligence systems in the Republic of Kazakhstan in the message of the president of the Republic of Kazakhstan dated October 5, 2018. News NAS RK, Ser. Geol. Tech. Sci. **6**, 47–54 (2018)
22. Schuster, J.A.: The scientific revolution. In: Companion to the History of Modern Science, pp. 217–242. Routledge, London (2006)
23. Suleimenov, I., Gabrielyan, O., Vitulyova, Y.: Dialectics of scientific revolutions from the point of view of innovations theory. Wisdom. **24**(4), 25–35 (2022). https://doi.org/10.24234/wisdom.v24i4.913
24. Brusentsov, N.P., Ramil Alvarez, J.: Ternary computers: the Setun and the Setun 70. In: Impagliazzo, J., Proydakov, E. (eds.) Perspectives on Soviet and Russian Computing, vol. 357, pp. 74–80. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22816-2_10
25. Suleimenov, I.E., Bakirov, A.S., Matrassulova, D.K.: A technique for analyzing neural networks in terms of ternary logic. J. Theor. Appl. Inf. Technol. **99**(11), 2537–2553 (2021)
26. Suleimenov, I., Gabrielyan, O., Kopishev, E., Kadyrzhan, A., Bakirov, A., Vitulyova, Y.: Advanced applications of polymer hydrogels in electronics and signal processing. Gels. **10**, 715 (2024). https://doi.org/10.3390/gels10110715
27. Ayan, Z., Alimzhan, B., Olga, M., Timur, Z., Toktalyk, Z.: Quality of Service Management in Telecommunication Network Using Machine Learning Technique. Indones. J. Electr. Eng. Comput. Sci. **32**(2), 1022–1030 (2023)

28. Suleimenov, I.E., Vitulyova, Y.S., Kabdushev, S.B., Bakirov, A.S.: Improving the efficiency of using multivalued logic tools. Sci. Rep. **13**(1), 1108 (2023)
29. Suleimenov, I.E., Vitulyova, Y.S., Kabdushev, S.B., Bakirov, A.S.: Improving the efficiency of using multivalued logic tools: application of algebraic rings. Sci. Rep. **13**(1), 22021 (2023)

Optimization of Winemaking Processes Using Artificial Intelligence Methods

Nadezhda Gallini[1], Anatoliy Kazak[1(✉)], Yuriy Grishin[2], Marina Rudenko[1], and Victor Gallini[1]

[1] V.I. Vernadsky Crimean Federal University, Simferopol 295007, Russia
kazak_a@mail.ru

[2] All-Russian National Research Institute of Viticulture and Winemaking" Magarach"RAS", 298600 Yalta, Russia

Abstract. This study presents a web-based intelligent system for modeling and optimizing winemaking processes with a focus on phenolic composition and antioxidant activity. The platform integrates data ingestion, preprocessing, empirical baselines and machine-learning modules, and an optimization layer that maps process settings to functional outcomes. Using curated datasets of white grape materials, the system predicts antioxidant activity from phenolic metrics and controllable technological factors and generates recommendations for process parameters. Cross-validated evaluation yields an R2 of 0.735 with MAE 0.061 and RMSE 0.073, indicating reliable explanatory power with low absolute error. Model interpretation highlights total phenols as the dominant positive predictor, with additional contributions from catechin, gallic acid, enzymatic maceration and contact time, and a negative association with pH. The system provides interactive visualizations and exportable reports, enabling rapid scenario testing and decision support for both production and research settings.

Keywords: Artificial intelligence · Winemaking · Process optimization · Phenolic composition · Antioxidant activity · Decision support

1 Introduction

The modern winemaking industry faces the challenge of simultaneously ensuring high product quality, compliance with safety standards, and environmental sustainability. Traditional approaches to selecting technological parameters are largely based on empirical data and the long-term expertise of specialists, which limits adaptability to changing production conditions and variability of raw materials. Meanwhile, the development of digital technologies and artificial intelligence methods opens new opportunities for comprehensive optimization of grape processing. Particular attention is paid to phenolic compounds, which determine the antioxidant properties of wine and affect its organoleptic profile. Building systems capable of predicting the behavior of these compounds under different processing regimes represents a strategically important direction in digital winemaking and sustainable food technologies.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 123–130, 2027.
https://doi.org/10.1007/978-3-032-27834-0_12

A review of recent studies demonstrates a consistent scientific interest in predicting antioxidant activity of plant materials and developing computational tools for assessing their biological value. Priyandoko et al. [1] and Widowati et al. [10] confirmed the high antioxidant activity of green tea extracts, soybeans, and isoflavones, highlighting the role of phenolic compounds as key predictors of biological activity. Hayati et al. [9] and Gulo et al. [5] expanded the range of investigated materials, demonstrating the contribution of flavonoids and polyphenols to the antioxidant potential of leaf and citrus extracts.

Another important area of research is the study of the antioxidant properties of non-traditional plant raw materials. Manin et al. [6] demonstrated the potential of tomato seeds, while Mehdi et al. [7] developed approaches to optimize extraction and enhance the antioxidant and antifungal activity of lavender essential oil. These studies emphasize the need to optimize technological regimes for the extraction and preservation of biologically active compounds.

Digital and computational approaches are also gaining momentum. An example of the effective application of machine learning methods in oenology is the work of Ramamoorthy et al. [8]. They applied random forest algorithms to predict the quality of the wine. In a study published by Madiyar and co-authors [3], the possibilities of using nanostructured materials with pronounced antioxidant characteristics were demonstrated. The research results open up new perspectives for the integration of biotechnological and digital methodological approaches in scientific and applied fields.

Special attention is paid to the integration of artificial intelligence into the practice of winemaking. De La Cruz Rojas et al. [2] proposed a methodology for predicting wine turbidity and optimizing clarifier dosage, while Izquierdo-Bueno et al. [4, 11] developed an intelligent decision support system aimed at managing risks and improving product quality. These studies underline the growing relevance of transitioning from localized analytical tasks to comprehensive intelligent systems capable of optimizing winemaking processes in real time.

Thus, the literature review shows significant progress both in studying antioxidant properties of raw materials and in developing predictive algorithms and decision-support systems for the wine industry. However, the comprehensive optimization of technological processes, taking into account the dynamics of phenolic compounds and antioxidant activity, remains insufficiently studied. This gap defines the objective of the present research.

2 Materials and Methods

The developed web-based system was implemented in Python using libraries for statistical analysis, machine learning, and data visualization. The system architecture incorporates a module for uploading experimental data on concentrations of bioactive compounds, a regression modeling module for predicting antioxidant activity, and an optimization module that determines technological parameters leading to maximum levels of flavan-3-ols, hydroxybenzoic acids, flavones, and other valuable phenolic compounds. Synthetic data generation is integrated to expand the training set, thereby improving algorithm robustness. The web interface supports interactive visualizations in the form of charts and tables and automatically generates recommendations for selecting optimal technological conditions.

The algorithm begins with the initialization of the working environment and the construction of a baseline model of wine composition, where reference values are set for key classes of phenolic compounds and the initial antioxidant activity index. The user uploads or edits the parameter table, after which the data are normalized through harmonization of column headers and measurement units to ensure compatibility with the system's internal nomenclature and to avoid errors when working with external datasets. Next, the user selects a processing scheme, such as classic white winemaking without skin contact, prefermentative maceration, or fermentation on skins and stems, and defines process parameters including contact time, temperature, addition of pectolytic enzymes, sulfur dioxide dosage, must acidity, ethanol presence during fermentation, stem fraction in the mash, and proportion of press wine in the blend. Based on the chosen scheme, empirical coefficients reflecting structural differences in extraction and transformation of phenolic compounds are applied to the baseline values, followed by the calculation of parameterized scaling factors modeling the kinetics of extraction and stabilization of phenolic compounds under the influence of time, temperature, enzymatic treatment, sulfur dioxide, acidity, ethanol, stem fraction, and press fraction. The resulting scaling multipliers are convolved with baseline values for each phenolic class to generate a predicted compositional profile, while antioxidant activity is calculated as a weighted function of phenolic metrics, taking into account nonlinear saturation effects for total monomeric and polymeric fractions.

In parallel, fold-change values relative to the baseline matrix are computed to interpret process effects quantitatively, with the results presented in tabular and graphical forms and exportable to standard reporting formats. When an experimental dataset is available, the system performs validation across all samples, automatically comparing observed and predicted values for antioxidant activity and phenolic classes, calculating mean absolute error, root mean square error, and Pearson correlation, and generating a quality summary for model performance assessment. In interactive mode, the user can retrieve baseline values from a reference technological scheme, substitute parameters directly from dataset rows for rapid recalculations, and visually compare predictions with experimental data. This functionality turns the algorithm into a tool for real-time process design and hypothesis testing under production conditions.

The workflow of the web-based system is illustrated in Fig. 1. At the initial stage, the system performs data input and normalization, including phenolic composition and technological parameters. It then offers two modes: applying built-in empirical coefficients or building an optimization model based on machine learning methods. In the first case, calculations are carried out using predefined equations, while in the second, the system trains a model on experimental data and determines optimal process parameters. The result is a predicted phenolic profile and antioxidant activity, followed by model validation using the coefficient of determination (R2), mean absolute error (MAE), and root mean square error (RMSE). At the final stage, the system generates visualizations of parameter dynamics, provides recommendations for technological regimes, and exports results in graphical and tabular formats. Learning was performed with linear regression using one-hot encoding of categorical factors and five-fold cross-validation with shuffle (random_state = 43). Metrics reported are R2, MAE, and RMSE.

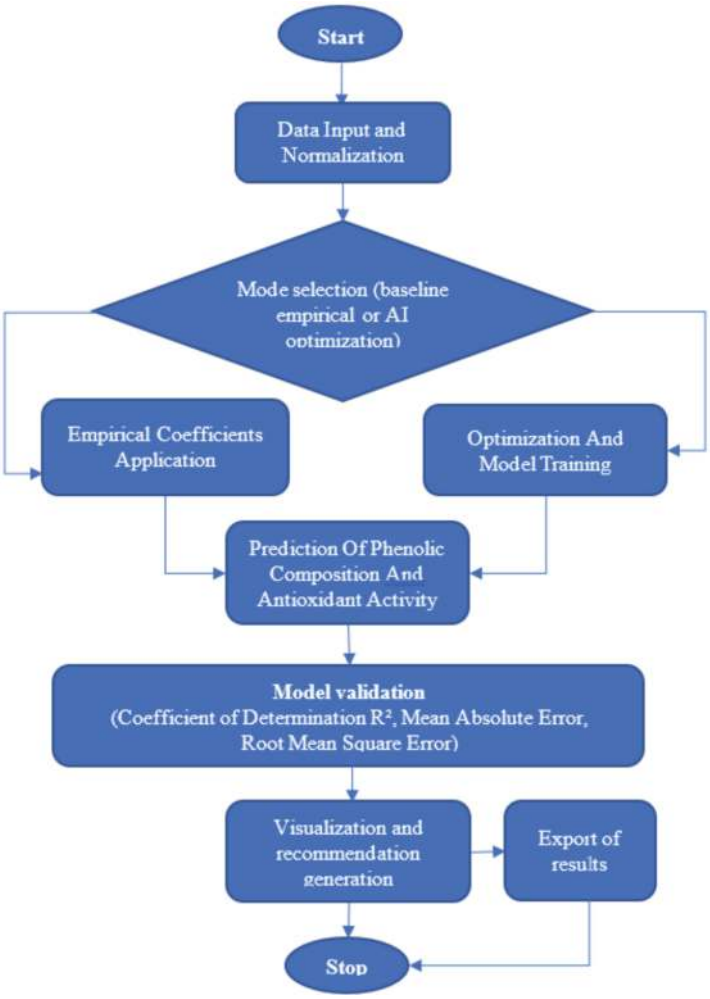


Fig. 1. Algorithm of the web-based system for optimization of winemaking technologies using artificial intelligence.

3 Results

The evaluation of the proposed web-based system on a curated dataset of winemaking materials demonstrated stable predictive performance. Cross-validated predictions of antioxidant activity achieved a coefficient of determination of 0.735 with a mean absolute error of 0.061 and a root-mean-square error of 0.073, indicating that the model explains about three quarters of the variance while retaining low absolute error across the range of measurements (Fig. 2).

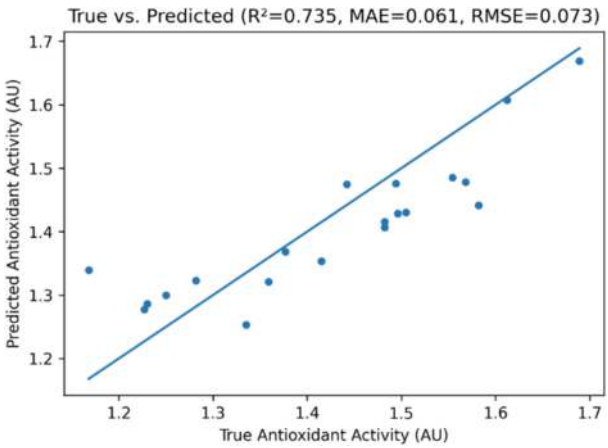


Fig. 2. Algorithm of the web-based system for optimization of winemaking technologies using artificial intelligence.

The residual distribution was approximately unimodal and centered close to zero with a mild positive tail and no apparent heteroscedasticity, which is consistent with a well-calibrated linear specification under the present data regime (Fig. 3). Inspection of the fitted coefficients revealed that the total phenolic content was the dominant positive predictor, while catechin and gallic acid also contributed positively to the antioxidant response. The inclusion of the processing context showed that enzymatic treatment, longer contact time, and moderately elevated temperature were associated with higher predicted antioxidant activity, whereas higher pH exhibited a negative association.

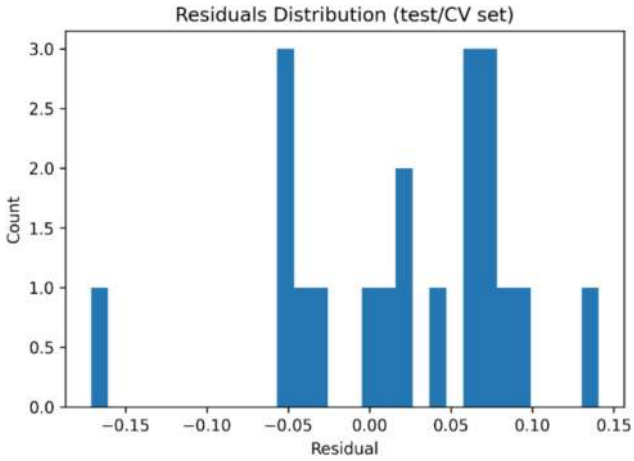


Fig. 3. Distribution of residuals for the test/CV predictions showing near-zero mean and no visible heteroscedasticity.

The indicator for processing on skins with stems had a positive contribution, reflecting the expected enhancement of phenolic extraction under this regime. Two regional indicators displayed small negative coefficients, suggesting site-specific differences in composition that are not fully captured by the measured phenolic variables. A negative coefficient for the flavan-3-ols aggregate was observed, which most likely reflects sign reversal under multicollinearity with total phenolics rather than a true inhibitory effect. The full set of standardized coefficients and their relative magnitudes are summarized in Fig. 4.

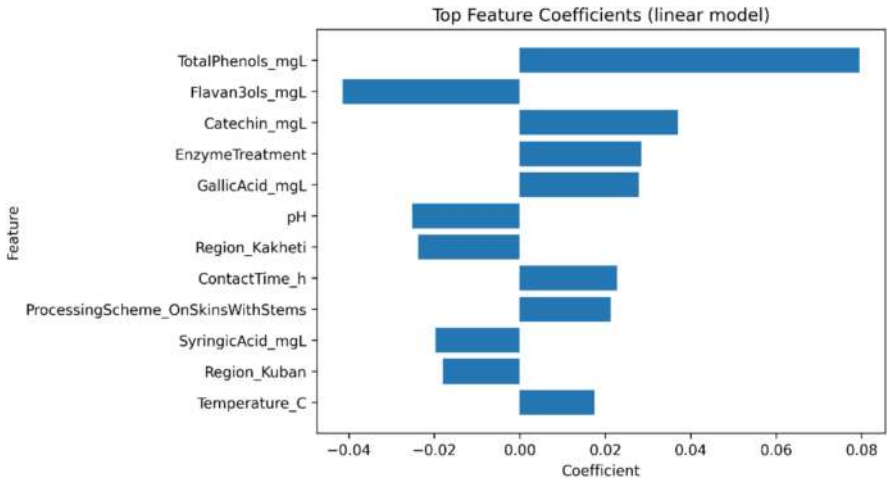


Fig. 4. Linear coefficients ranked by absolute magnitude; total phenols is the dominant positive predictor, pH is negatively associated.

4 Discussion

The introduction of an intelligent system has demonstrated significant advantages over traditional methods of process optimization. The use of artificial intelligence has significantly reduced the time needed to select processing modes and reduced dependence on the subjective opinions of experts. In addition, the system provides personalized recommendations for specific grape varieties, which increases the reproducibility and predictability of the final product. Future research should focus on expanding the database to include materials from various geographical regions and grape varieties, as well as integrating more complex models such as neural networks and ensemble methods to improve the accuracy and reliability of forecasts.

The results confirm that the phenolic composition and controlled technological parameters together serve as an informative basis for predicting antioxidant activity in oenological matrices. The predominance of total phenols among the predictors aligns with mechanistic expectations and prior reports linking global phenolic load to redox capacity. Positive effects of catechin and gallic acid further support the role of monomeric flavan-3-ols and hydroxybenzoic acids in driving the signal, while the negative pH association is consistent with the higher stability and extractability of phenolic forms under

more acidic conditions. The positive contributions from enzymatic maceration, contact time, and temperature capture known extraction kinetics and indicate that the system can translate process settings into compositional shifts with measurable functional consequences. The negative coefficient estimated for the summed flavan-3-ols should be interpreted cautiously, as diagnostic plots and pairwise correlations suggest substantial overlap with total phenolics. In such cases, coefficient shrinkage or partial least squares can stabilize signs without materially changing out-of-sample fit. Residual analysis does not reveal systematic bias across the target range, implying that the current linear specification is adequate for decision support within the observed domain. Extending the dataset with additional varieties, wider ranges of maceration intensity, and explicit measurement-method indicators is expected to improve generalization and reduce the influence of collinearity, particularly for families of phenolics that co-vary with total content. To mitigate coefficient instability under multicollinearity, ridge or partial least squares can be incorporated without sacrificing interpretability, and are planned for the next release of the system.

5 Conclusions

The developed intelligent web-based system has proven effective for modeling and optimizing winemaking processes. The results confirm its practical value in increasing antioxidant activity and the biological value of winemaking products. The system combines predictive accuracy with usability and can serve as a foundation for implementing digital solutions in the winemaking industry.

Software Availability

Digital web-calculator for tartaric acid stock prediction using AI, **RU 2025685854.**

References

1. Priyandoko, D., Pratiwi, S.D., Mulyono, A.R., Nugroho, A.K.: Antioxidant activity of green tea extract and Myricetin. In: Proceedings 2021 IEEE International Conference on Health, Instrumentation & Measurement, and Natural Sciences (InHeNce), Medan, Indonesia, pp. 1–5 (2021), https://doi.org/10.1109/InHeNce52833.2021.9537285
2. De La Cruz Rojas, E.M., Nuñez-Pérez, J., Lara-Fiallos, M., Pais-Chanfrau, J.-M., Espín-Valladares, R., DelaVega-Quintero, J.C.: An artificial intelligence approach for estimating the turbidity of artisanal wine and dosage of clarifying agents. Appl. Sci. **14**(11), 4416 (2024), https://doi.org/10.3390/app14114416
3. Madiyar, F., Aman, S., Fazel, M.A., Zhaksylyk, B.: Nomination for-electrosprayed thin films of multifunctional melanin nanoparticles for photoprotective and antioxidant properties. In: Proceedings 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, pp. 1–5 (2024), https://doi.org/10.1109/EMBC53108.2024.10781697
4. Izquierdo-Bueno, I., Moraga, J., Cantoral, J.M., Carbú, M., Garrido, C., González-Rodríguez, V.E.: Smart viniculture: applying artificial intelligence for improved winemaking and risk management. Appl. Sci. **14**, 10277 (2024). https://doi.org/10.3390/app142210277

5. Gulo, K.N., Suhartomi, Saragih, A.D., Raif, M.A., Ikhtiari, R.: Antioxidant activity of flavonoid compounds in ethanol and ethyl acetate extract from citrus Sinensis. In: Proceedings 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, pp. 1–6 (2021), https://doi.org/10.1109/AIMS52415.2021.9466078
6. Manin, L., Jelić, T.A., Merunka, D., Jurec, J., Muhvić, D., Valić,S.: Analysis of antioxidant power and application of cherry tomato seeds. In: Proceedings 2024 International Workshop on Quantum & Biomedical Applications, Technologies, and Sensors (Q-BATS), Durrës, Albania, pp. 66–70 (2024), https://doi.org/10.1109/Q-BATS63267.2024.10873877
7. Mehdi, L., Ayoune, A., Bouchema, D., Ouassila, L., Abdeslam-Hassen, M.: Optimization of extraction of essential oil from local lavender (Lavandula angustifolia) and evaluation of the antioxidant and antifungal activities. In: Proceedings 2022 13th International Renewable Energy Congress (IREC), Hammamet, Tunisia, pp. 1–5 (2022),https://doi.org/10.1109/IREC56325.2022.10002018
8. Ramamoorthy, M., Tarra, J.R., Kalidas, N., L.S., Karthick, V., Deepakkumar, K.: Prediction model for wine quality prediction using random forest. In: Proceedings 2024 International Conference on Advances in Computing, Communication and Materials (ICACCM), Dehradun, India, pp. 1–4 (2024), https://doi.org/10.1109/ICACCM61117.2024.11059098
9. Hayati, W., Raif, M.A., Ginting, C.N., Ikhtiari, R.: Antioxidant and wound healing potential of Persea Americana Mill. leaves extract. In: Proceedings 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, pp. 1–5 (2021), https://doi.org/10.1109/AIMS52415.2021.9466076
10. Widowati, W., Ikhtiari, R., Mustofa, S.A., Nur, K.M.: Antioxidant properties of Soybean (Glycine max L.) extract and Isoflavone. In: Proceedings 2021 IEEE International Conference on Health, Instrumentation & Measurement, and Natural Sciences (InHeNce), Medan, Indonesia, pp. 1–6 (2021), https://doi.org/10.1109/InHeNce52833.2021.9537282
11. Vinodha, D., Buvana, M.J., Rajalakshmi, S., Anita, E.A.M., Lapina, M.: Framework based on IoT, AI, and blockchain for smart access to government agricultural schemes, the future of agriculture: IoT, AI and Blockchain Technology for Sustainable Farming, pp 142–163 (2024), https://doi.org/10.2174/9789815274349124010011

Application of Artificial Intelligence Methods in a Program for Predicting the Antioxidant Activity of Winemaking Materials

Nadezhda Gallini[1], Anatoliy Kazak[1(✉)], Marina Rudenko[1], Yuriy Grishin[2], Anisa Atik[1], and Victor Gallini[1]

[1] V.I. Vernadsky Crimean Federal University, Simferopol 295007, Russia
kazak_a@mail.ru

[2] "All-Russian National Research Institute of Viticulture and Winemaking" Magarach"RAS", 298600 Yalta, Russia

Abstract. The paper presents a software solution designed to predict the antioxidant activity (AA) of winemaking materials based on their phenolic composition. The program relies on regression models that establish quantitative relationships between the concentration of phenolic compounds and the level of antioxidant activity. By entering chemical analysis data, users can quickly and accurately assess the antioxidant properties of wine. This functionality is important for quality control, the development of functional products, and the optimization of technological processes. The program reduces the time and costs of laboratory analyses while providing a reliable tool for assessing the biological value of winemaking materials. It can be used by wineries, research institutions, and quality control laboratories.

Keywords: Artificial intelligence · Antioxidant activity · Winemaking materials · Phenolic compounds · Regression analysis · Quality control

1 Introduction

Antioxidant activity is one of the key indicators of the biological and functional value of winemaking materials. Phenolic compounds found in grapes and wine products play a crucial role in shaping antioxidant potential and influencing many organoleptic properties. Over the past decades, considerable attention has been paid to studying the antioxidant properties of plant extracts, as reflected in numerous publications. Priyandoko et al. (2021) demonstrated high antioxidant activity in green tea extract and myricetin, while Gulo et al. (2021) investigated the antioxidant properties of flavonoids in citrus extracts. Similar results were reported by Hayati et al. (2021) for avocado leaf extract and by Widowati et al. (2021) for soybean isoflavones, confirming the importance of phenolic compounds as key determinants of antioxidant activity.

Other studies have examined a wide range of plant-based materials. Johnson et al. (2024) develops prediction of antioxidant capacity in faba bean from individual phenolic

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 131–135, 2027.
https://doi.org/10.1007/978-3-032-27834-0_13

constituents. Kazak et al. (2022) analyzed the use of machine learning for comparative analysis of amperometric and chemiluminescent methods for determining antioxidant activity and determining the phenolic profile of wines.

2 Materials and Methods

In this study, a regression analysis of a dataset containing the characteristics of the phenolic composition and antioxidant activity of wine raw materials was performed. As a result, predictive models were created using optimization methods using machine learning algorithms. A program with an interface has been developed for the automatic prediction of AOA based on the entered data on the mass concentration of phenolic compounds. The AOA forecast can be created in two ways. In the first case, predefined regression equations of the form $y = b_0 + \Sigma(b_i \cdot x_i)$ are used, where y is the antioxidant activity, and x is the mass concentration of phenolic compounds, such as the total amount of phenols and individual acids. The second way is to train the model on its own dataset of phenolic composition and AOA. Data is automatically distributed to training and test sets, and a regression model is built. The accuracy indicators of the model are the mean absolute error (MAE), the RMS error (RMSE), and the coefficient of determination (R^2). As a result of the training, forecasts of the AOA of new datasets are created. The program allows you to build various graphs that visualize AOA forecasts.

The implementation was performed using Python libraries, including scikit-learn (for regression models and cross-validation), pandas and numpy (for data processing), plotly (for visualization), and streamlit (for the user interface). Figure 1 shows the algorithm of the program for predicting antioxidant activity.

The system functions in two modes: using empirical regression equations and training a model on experimental data. In the first case, concentrations of phenolic compounds are entered, and the antioxidant activity value is calculated. In the second case, the dataset is uploaded, the model is constructed and evaluated, and predictions are generated for new samples. In both cases, the program provides predictions and visualization of the results.

Testing was performed on a dataset of winemaking materials with known experimental values of antioxidant activity. Comparison of predicted and measured results showed a high degree of agreement. Correlation coefficients exceeded 0.9, confirming the reliability of the constructed models. These findings demonstrate that the program can be used for rapid assessment of the antioxidant properties of winemaking products and for significantly reducing the costs of additional laboratory analyses (Fig. 1).

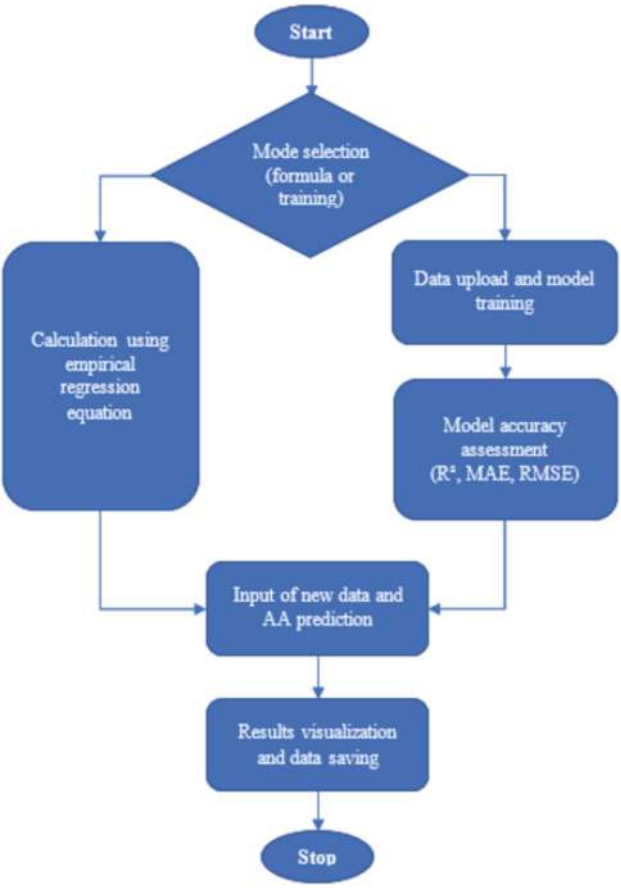


Fig. 1. Program algorithm.

3 Results and Discussion

Testing was carried out on a dataset of winemaking materials with known experimental values of antioxidant activity. Comparison of predicted and measured data showed a high degree of agreement, with correlation coefficients exceeding 0.9. This confirms the reliability of the constructed models and demonstrates the potential of the program for rapid assessment of antioxidant properties, significantly reducing the need for additional laboratory testing.

The program processes various data efficiently, but its accuracy increases with an increase in the size of the analyzed datasets, which include the characteristics of wine materials, and the use of tools such as, for example, ensemble methods. Figure 2 visualizes the distribution of dataset errors relative to the true values of AOA. The errors are distributed in a narrow range and symmetrically around zero, which shows the good quality of the model.

These visualizations highlight that the model not only provides high numerical accuracy ($R2 = 0.9567$, $MAE = 0.1054$, $RMSE = 0.1240$) but also demonstrates consistency

across different samples. This further supports the suitability of the proposed system for practical applications in winemaking quality control (Fig. 2).

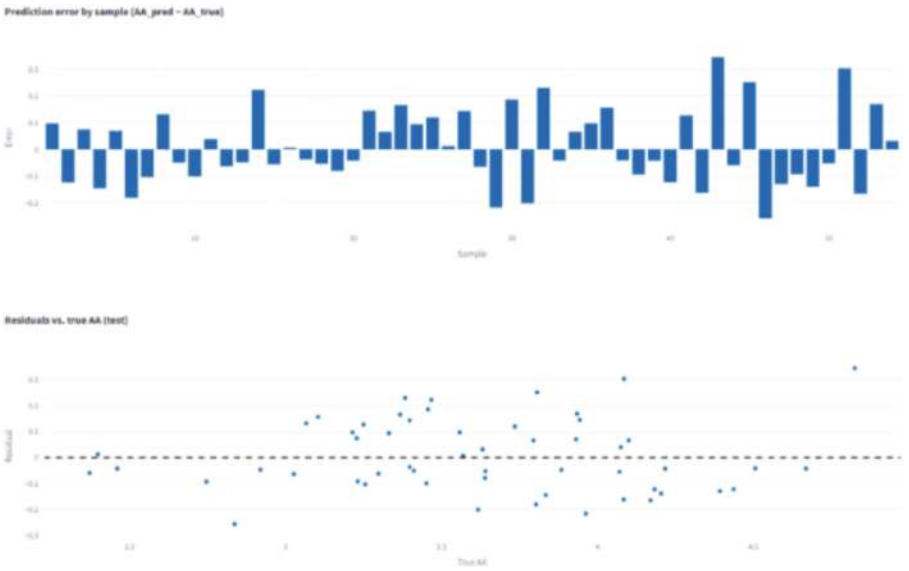


Fig. 2. Error distribution analysis.

4 Discussion

The program created in this study allows us to make high-quality and rapid forecasts of the relationship between AOA and the mass concentration of phenolic substances. Increasing the size of datasets and using ensemble methods makes it possible to further improve the reliability of forecasting.

5 Conclusions

In this study, a regression analysis of a dataset containing the characteristics of the phenolic composition and antioxidant activity of wine raw materials was performed. As a result, predictive models were created using optimization methods using machine learning algorithms. A program with an interface has been developed for the automatic prediction of AOA based on the entered data on the mass concentration of phenolic compounds. Thanks to the use of such tools, it is possible to improve production efficiency and improve the quality of wine.

References

Priyandoko, D. et al.: Antioxidant activity of green tea extract and Myricetin. 2021 IEEE International Conference on Health, Instrumentation & Measurement, and Natural Sciences (InHeNce), Medan, Indonesia, pp. 1–5 (2021), https://doi.org/10.1109/InHeNce52833.2021.9537285

Madiyar, F. et al.: Nomination for-electrosprayed thin films of multifunctional melanin nanoparticles for photoprotective and antioxidant properties. 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, pp. 1–5 (2024), https://doi.org/10.1109/EMBC53108.2024.10781697

Gulo, K.N., Suhartomi, Saragih, A.D., Raif, M.A., Ikhtiari, R.: Antioxidant activity of flavonoid compounds in ethanol and ethyl acetate extract from citrus Sinensis. 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, pp. 1–6 (2021), https://doi.org/10.1109/AIMS52415.2021.9466078

Manin, L., et al.: International workshop on quantum & biomedical applications, technologies, and sensors (Q-BATS). Durrës, Albania **2024**, 66–70 (2024). https://doi.org/10.1109/Q-BATS63267.2024.10873877

Mehdi, L., Ayoune, A., Bouchema, D., Ouassila, L., Abdeslam-Hassen, M.:Optimization of extraction of essential oil from local Lavender (lavandula angustifolia) and evaluation of the antioxidant and antifungal activities. 2022 13th International Renewable Energy Congress (IREC), Hammamet, Tunisia, pp. 1–5 (2022), https://doi.org/10.1109/IREC56325.2022.10002018

Hayati, W., Raif, M.A., Ginting, C.N., Ikhtiari, R.: Antioxidant and wound healing potential of Persea Americana Mill. Leaves extract. 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, pp. 1–5 (2021), https://doi.org/10.1109/AIMS52415.2021.9466076

Johnson, J.B., Kazak, A., Gallini, N., et al.: Prediction of antioxidant capacity in faba bean from individual phenolic constituents. Chem. Pap. **78**, 4285–4294 (2024). https://doi.org/10.1007/s11696-024-03389-3

Kazak, A., et al.: The use of machine learning for comparative analysis of amperometric and chemiluminescent methods for determining antioxidant activity and determining the phenolic profile of wines. Appl. Syst. Innov. **5**, 104 (2022). https://doi.org/10.3390/asi5050104

Composition of Differential Privacy Mechanisms in Iterative Machine Learning: A Comparative Study

Saveliy Gordeev[1(✉)], Anton Sinitsyn[2], and Maria Lapina[1]

[1] North-Caucasus Federal University, 2 Kulakova Ave, Stavropol 355029, Russia
gordeeff.sava@mail.ru
[2] Russian Academy of Sciences, 109004 Moscow, Russia
antony@email.su

Abstract. This paper deals with theoretical and practical aspects of differential privacy application in iterative machine learning algorithms. The main attention is paid to the analysis of privacy composition mechanisms and their influence on the quality and stability of trained models. A comparative analysis of two basic mechanisms, Gaussian and Laplace, providing different levels of privacy and accuracy, is carried out. The basic methods of composition are described and their influence on the dynamics of accumulation of the leakage parameter ε is investigated. The experimental part is performed in Python using PyTorch and Opacus libraries on the MNIST dataset. The work emphasises the importance of correct choice of privacy mechanism and composition method in building secure machine learning models. The results obtained can be used to develop effective private learning strategies that provide an optimal compromise between data protection and practical usefulness of models.

Keywords: Diffential privacy · Machine learning · Data privacy · Information security · Python

1 Introduction

Most modern machine learning (ML) [7,10,18,21] algorithms are actively used to process large amounts of data, often such data carry sensitive information. However, the use of such data generates a high risk of important data leakage. Standard information security methods are proving insufficient, as even aggregated data can be recovered with a high degree of accuracy due to privacy attacks. In response to this problem, the concept of differential privacy (DP) [5] has become widespread in recent years. This concept characterises a formal approach that provides robust mathematical methods for protecting individual data. The main idea of differential privacy is to add controlled noise to the model training process, which makes it impossible to determine the influence of an individual object on the final result. Differential privacy is of particular

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 136–148, 2027.
https://doi.org/10.1007/978-3-032-27834-0_14

interest in the context of iterative machine learning algorithms [1]. In machine learning models based on the iterative approach, privacy is ensured at the level of each iteration, which leads to the so-called composition effect. The essence of this effect is the gradual accumulation of the leakage parameter ε as the model iterates. Both the final accuracy of the model and the actual level of data protection directly depend on the choice of the composition mechanism and the way privacy is taken into account. The problem of composition is one of the key problems in the theory of differential privacy, because the way it is taken into account directly affects both the degree of privacy and the accuracy of the trained model. Various approaches to estimating the accumulated level of privacy such as improved composition, method of moments, as well as more recent methods based on Renyi Differential Privacy (RDP) [12] and zero-Concentrated Differential Privacy (Zero-Concentrated Differential Privacy) have been developed. (zCDP) [3], perform the approximation of ε growth differently and provide a different balance between reliability and accuracy of learning. In addition, the efficiency and robustness of algorithms with differential privacy are significantly affected by the parameters of the iterative process: the size of the batches, the learning step, the frequency of noise addition, and the form of the mechanism used . Incorrect choice of these parameters can lead either to loss of privacy or to a sharp deterioration of model accuracy due to strong noise in the gradients. The relevance of the study stems from the need for a thorough understanding of how different differential privacy mechanisms and their composition methods interact in conjunction with iterative machine learning methods. Current practice shows that the trade-off between privacy and accuracy remains one of the most challenging problems in the field of Privacy-Preserving Machine Learning (PPML) [2,6,10]. The aim of this study is to perform a comparative analysis of the compositional properties of various differential privacy mechanisms in iterative machine learning algorithms, and to select an optimal tool for different ML models. The results of the analysis allow us to identify the peculiarities of the interaction between different different differential privacy mechanisms and iterative optimisation processes and to determine the optimal combinations of parameters that provide a balance between the level of privacy and the quality of the model.

2 How Differential Privacy Works

Unlike traditional approaches based on encryption or anonymisation, DP does not hide the data itself directly, but makes it impossible to reliably determine the participation of a particular element in the training set. In other words, it makes the algorithm resistant to so-called inference attacks, in which an attacker tries to determine from the result of a model whether a particular item was present in the training sample. Thanks to DP, data leaks by means of such attacks are avoided due to the DP construction mechanism.

Let $\mathcal{D}$ be the underlying data space, and let two data sets $D, D' \in \mathcal{D}$ be called adjacent if they differ in only one element. This means that exactly one

entry in the set D' is replaced or deleted from the set D. This definition allows us to formalise the minimal possible change in the dataset associated with the participation of a single individual. Next, we consider a stochastic mechanism (or algorithm) $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ that takes a data set as input and produces some result, often expressed in a trained ML model, the mean value of a feature or statistical object of interest.

The meaning of differential privacy is that the result of the $\mathcal{M}$ mechanism is practically independent of the presence or absence of some individual element in the data. Formally, this is expressed by restricting the ratio of the probabilities of the outputs $\mathcal{M}(D)$ and $\mathcal{M}(D')$. For any two adjacent sets D and D' and for any possible outcome $S \subseteq \mathcal{R}$, the condition must hold

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta. \tag{1}$$

The parameters ε and δ quantify the level of privacy. The smaller the value of ε, the stronger the privacy guarantees, since changing one record has almost no effect on the distribution of algorithm outputs. The δ parameter allows for a small probability of privacy violation, e.g., due to numerical or stochastic features of the mechanism. In practical scenarios, δ is usually chosen $\leq 10^{-5}$ or less, and ε tends to be kept in the range from 0.1 to 5, depending on the requirements for accuracy and level of protection, as well as the peculiarities of building the structure of machine learning algorithms.

Intuitively, one can imagine that if the $\mathcal{M}$ mechanism is (ε, δ) differentially private, then an observer with access only to the result of $\mathcal{M}(D)$ will not be able to determine with high probability whether a particular object was included in the dataset or not. This approach gives universal mathematical guarantees that are independent of external factors and assumptions about the adversary.

From a practical point of view, achieving differential privacy requires introducing controlled noise into the computation. Such noise is added either at the input data level, often in intermediate layers or in gradients during model training. The main challenge in doing so is to strike a balance: the noise must be large enough to protect the data, but not so large that it destroys useful patterns and results in a loss of model accuracy. It is this trade-off between privacy and the quality of the resulting model that is the central problem of the whole theoretical component of differential privacy.

Next, let us consider the basic mathematical mechanisms underlying the construction of learning algorithms using differential privacy.

3 Mechanisms of Composition and Their Comparative Analysis

The most important feature of DP when applied to iterative machine learning algorithms is that the privacy mechanism is triggered multiple times at each iteration of updating the model parameters. Each application of the mechanism contributes to the overall privacy level expressed by the parameters (ε, δ). Over

time, there is an increase in these parameters, which leads to the need to accurately account for the composition of all iterations. That is why the theory of composition is fundamental in the task of constructing DP-based algorithms.

Suppose we have a sequence of mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_T$, each of which provides a $(\varepsilon_t, \delta_t)$ measure of DP. The combined application of these mechanisms then forms a composite mechanism $\mathcal{M}_{1:T}$, for which the aggregate privacy parameters $(\varepsilon_{\text{tot}}, \delta_{\text{tot}})$ depend on the way the composition is organised. The naive approach involves simple summation:

$$\varepsilon_{\text{tot}} = \sum_{t=1}^{T} \varepsilon_t, \qquad .\delta_{\text{tot}} = \sum_{t=1}^{T} \delta_t. \tag{2}$$

However, this estimation is highly unreliable and overestimates the privacy loss rate, especially when the number of iterations is large. Therefore, improved composition methods have been proposed that produce denser and more realistic estimates.

One such method is improved composition, according to which, if each step of the algorithm provides (ε, δ_0) level of DP measures, then the entire process of T iterations satisfies $(\varepsilon', T\delta_0 + \delta')$ degree of DP, where

$$\varepsilon' = \sqrt{2T \ln(1/\delta')}\, \varepsilon + T\varepsilon(e^{\varepsilon} - 1). \tag{3}$$

This formula shows that for moderate values of ε, the cumulative leakage of ε' does not grow linearly, but roughly proportional to $\sqrt{T}$, which significantly improves the balance between privacy and accuracy of the model.

Modern compositional approaches are often based on more sophisticated mathematical methods. These methods allow us to describe the distribution of privacy losses at each step and aggregate them in an analytically convenient form. For example, RDP has the property of additivity in the number of iterations:

$$\varepsilon_{\alpha}^{\text{total}} = \sum_{t=1}^{T} \varepsilon_{\alpha,t}. \tag{4}$$

which makes it particularly useful for analysing differential-private stochastic gradient descent (DP-SGD) [19]. After summing over iterations, the RDP can be converted to the standard value of the exponent (ε, δ) via the ratio

$$\varepsilon = \varepsilon_{\alpha} + \frac{\ln(1/\delta)}{\alpha - 1}, \tag{5}$$

here the parameter α is chosen to minimise the final value of ε. Similarly, zCDP is expressed through the ρ parameter, which is also linearly summed during composition and then converted to (ε, δ) values using the formula

$$\varepsilon = \rho + 2\sqrt{\rho \ln(1/\delta)}. \tag{6}$$

Thus, the choice of composition mechanism directly affects the rate of growth of the indices and hence the resulting trade-off between model privacy and accuracy. In practical implementations of DP-SGD, the moment-based and RDP-based methods perform best, as they produce denser estimates at reasonable computational complexity. Next, let us take a closer look at the two basic mechanisms underlying most approaches, namely the Gaussian [4] mechanism and the Laplace [15] mechanism.

3.1 Gaussian Mechanism

The Gaussian mechanism is one of the most common tools for realising differential privacy in machine learning. Its idea is to add normally distributed noise to the result of a function whose sensitivity is limited. Let the function $f : \mathcal{D}R^d$ has ℓ_2-sensitivity Δ_2 defined as

$$\Delta_2 = \max_{D,D'} \|f(D) - f(D')\|_2, \tag{7}$$

where D and D' are contiguous datasets differing by one element. Then the Gaussian mechanism is defined by the expression:

$$\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2 I_d), \tag{8}$$

where σ is the noise scale parameter. For a given mechanism, it can be shown that it provides (ε, δ) degree DP if

$$\sigma \geq \frac{\Delta_2 \sqrt{2 \ln(1.25/\delta)}}{\varepsilon} \tag{9}$$

Thus, the larger the σ, the more the influence of the individual record is masked, but the higher the noise level and thus the potential decrease in model accuracy. The Gaussian mechanism has a number of advantages due to the implementation of the mechanism in RDP and zCDP, and also lends itself well to composition. In turn, the RDP implementation of the Gaussian mechanism has a simple form:

$$\varepsilon_\alpha^{\text{Gauss}} = \frac{\alpha \Delta_2^2}{2\sigma^2} \tag{10}$$

which makes it easy to estimate the accumulated noise level at a large number of iterations. From a practical point of view, the Gaussian mechanism is robust to error accumulation and has smooth aggregation properties, which makes it the preferred choice for problems with high dimensionality of model parameters.

3.2 Laplace Mechanism

The Laplace mechanism is an alternative approach to privacy based on adding noise that is distributed according to Laplace's law, namely a bilateral exponential distribution. For the function $f : \mathcal{D} \to R^d$ with ℓ_1-sensitivity Δ_1:

$$\Delta_1 = \max_{D,D'} \|f(D) - f(D')\|_1, \tag{11}$$

Laplace's mechanism is given by the formula:

$$\mathcal{M}(D) = f(D) + (Y_1, \ldots, Y_d), \qquad Y_i \sim \text{Laplace}\left(0, \frac{\Delta_1}{\varepsilon}\right) \tag{12}$$

With this added noise, the mechanism satisfies the strict $(\varepsilon, 0)$ degree of DP privacy. Laplace noise is characterised by sharper vertices of the distribution, which makes it sensitive to outliers, but still achieves a strong privacy guarantee without the δ parameter.

In practical scenarios, the Laplace mechanism is less frequently used, but remains an important theoretical tool for privacy analysis, as it demonstrates the boundaries between strict and approximate differential privacy. Its use is appropriate when guaranteed protection without probabilistic assumptions is required and the amount of data and number of iterations are limited.

Thus, the presented mechanisms are the basis for most differential privacy implementations, but the choice between them is determined by the specifics of the problem.

Further stages of our study, we will consider both classical methods of stochastic gradient descent (SGD) and their differential-private modifications (DP-SGD), as well as the feasibility of using these mechanisms in more complex ML algorithms. Different privacy-aware schemes will be investigated to evaluate the effectiveness of the compositional properties.

4 Experimental Part

The aim of the experimental part is to test the properties of different differential privacy mechanisms under iterative learning conditions. The main focus is on comparing the behaviour of the models when using different ways of adding noise, strategies of privacy consideration, and schemes of changing the learning parameters.

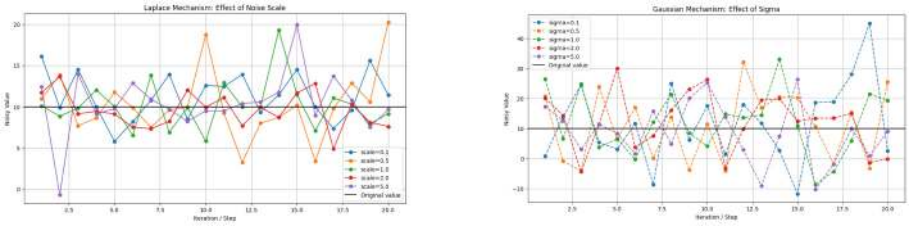
4.1 Noise Handling

The Python [17] programming language was chosen to implement the experiments using the PyTorch [14] library and the OPACUS [22] library providing the implementation (DP-SGD). The open image set MNIST [8] was used as input data.

To ensure privacy during model training, mechanisms were implemented to add random noise to the output values. A software implementation was carried out to add Laplace noise and Gaussian noise. Both mechanisms use the parameters ϵ and δ. During this analysis we obtained graphs (Fig. 1) showing the behaviour of the parameters depending on their value

The privacy parameter plot shows different dynamics of private leakage accumulation for the two mechanisms. Laplace noise, which provides more stringent guarantees, shows faster parameter growth for the same values of the noise

scale. At the same time, the Gaussian mechanism allows a more economical consumption of the privacy budget at the expense of relaxed guarantees. Next, the fundamental trade-off between the two mechanisms, reflected by the relationship between the proper level of confidentiality and the accuracy of the results (Fig. 2), was considered.



a) Gaussian mechanism b) Laplace mechanism

Fig. 1. Behaviour of different noise levels in machinery.

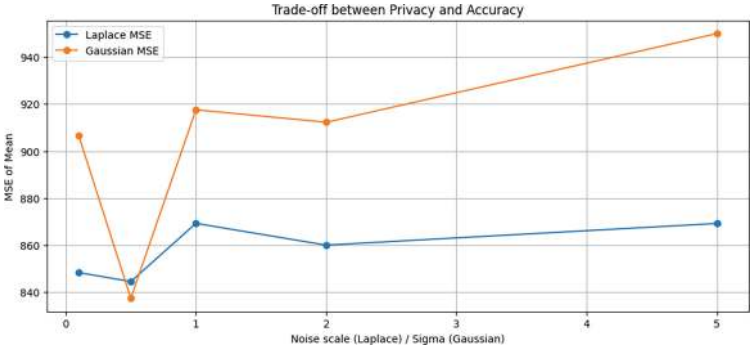


Fig. 2. Analysis of the trade-off between privacy and accuracy.

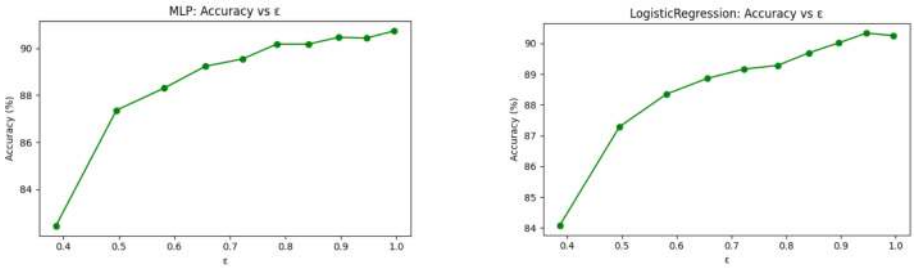
The difference in the behaviour of the curves highlights the importance of making an informed choice of protection mechanism depending on the requirements of a particular task. In situations where relaxed privacy guarantees are acceptable, the Gaussian mechanism may provide a more favourable balance between accuracy and privacy. However, in cases requiring the most stringent protection, the Laplace mechanism should be favoured, despite the greater reduction in accuracy.

4.2 Compositional and Parametric Analysis

Let us consider in more detail the impact of differential privacy on different neural network architectures and the dynamics of the learning process. The conducted research covers a number of machine learning models, which allows us to reveal fundamental patterns of interaction between the complexity of the architecture and DP requirements. Several types of ML models, namely logistic [11] and linear regression [20], as well as multilayer perceptron [16], were considered in the experiments.

In the context of this, let us highlight the properties of models of interest to us and consider a practical example of how models behave when working closely with DPs. The architecture of the multilayer perceptron (MLP) occupies an intermediate position, demonstrating a balanced behaviour. On the one hand, its fully-connected architecture allows to compensate quite effectively for the effects of noise by spreading errors over multiple connections, but on the other hand, the lack of specialised inductive biases makes it less efficient than more complex architectures on computer vision tasks at the same privacy levels. Logistic and linear regression, as the simplest models, show interesting properties in the DP context. Their robustness to noise appears to be the highest among the architectures considered, but due to the limited capacity of these models, achieving high absolute accuracy even at minimum noise levels is not guaranteed.

Thus, the main focus of the study will be on testing the compatibility of DP with ML models in order to maximise protection. First, let us consider the process of error accumulation during training in the presented models (Fig. 3):



a) MLP b) Logistic regression

Fig. 3. Noise accumulation during model training.

In the graphs we can trace the characteristic trajectories of error accumulation during training for two different model architectures. For the multilayer perceptron, we observe a non-linear dependence of accuracy on the ε value, where the initial rapid growth is followed by a gradual plateau. This indicates that the model effectively utilises the initial "privacy budget" but reaches its capacity limit when the constraints are further relaxed.

In the case of logistic regression, the dynamics shows a smoother approximation to a linear plot, which corresponds to the dependence of the absence of hidden layers and non-linear transformations prevents a significant improvement in accuracy even when the allowed privacy level is increased.

In the case of simple models, we observe a characteristic pattern in that the architecture of these models does not impose any particular constraints on the integration with DP. Consequently, it is necessary to analyse the most complex ML models and compare the performance. The architecture considered below is a CNN [13] convolutional neural network model with two convolutional layers followed by fully connected layers. This architecture is of particular interest to investigate in the context of differential privacy because it combines spatial data processing via convolutional operations and nonlinear transformations via activation functions. The convolutional layers of CNNs, with their properties of parametric exchange and spatial invariance, should theoretically exhibit increased robustness to noise added within DP mechanisms. However, the hierarchical structure of the network and the cascading nature of the transformations may also contribute to the accumulation and amplification of error through different layers. Thus, for further research, we will compile an exponentiated dependence for a number of parameters and analyse the results obtained. Initially, we trace the dependence between adaptive optimisers and compare optimizer performance in Table 1.

Table 1. Comparison of optimisers under different gradient cut-off parameters.

Optimizer	Clip	Accuracy, %	Loss
Adam	0.5	97.18	0.139
Adam	1.0	97.03	0.144
Adam	2.0	96.98	0.133
Adam	5.0	96.67	0.138
AdamW	0.5	97.21	0.142
AdamW	1.0	97.26	0.129
AdamW	2.0	97.25	0.124
AdamW	5.0	96.81	0.132
SGD	0.5	81.89	0.729
SGD	1.0	88.67	0.424
SGD	2.0	93.72	0.244
SGD	5.0	95.57	0.166

The study showed that optimisers such as Adam and AdamW demonstrate high robustness to noise due to their built-in adaptive learning mechanisms. Adam, which uses moving averages of the first and second moments of the gradients, effectively smoothes out fluctuations caused by the addition of DP noise. This

allows the learning dynamics to remain stable even under significant perturbations of the gradients. AdamW, in turn, further improves the regularisation process by separately controlling the decay weights and learning rate, which is particularly valuable under privacy budget constraints when each model parameter must be used to maximum efficiency.
Classical SGD, although theoretically transparent, exhibits significant sensitivity to the gradient cutoff parameter. Its performance strongly depends on the correct choice of this hyperparameter, since the error accumulation process directly depends on it.

A key aspect is the interaction between the optimiser and the gradient cutoff mechanism. Adaptive methods are able to compensate for the distortions introduced by the Clip operation by adapting individual learning steps for each parameter. This explains their superiority over SGD in differential privacy settings, where controlling the gradient norm is a mandatory requirement for privacy guarantees. Thus the influence of optimisers has a significant impact on the holistic performance of the ML model using DP.

We now consider the impact of various privacy composition [9] methods, which play a key role in correctly accounting for the cumulative loss of privacy when privacy mechanisms are repeatedly applied. In the context of neural network training, where each epoch represents many iterations of parameter updates, the choice of composition method directly affects the accuracy of the total privacy budget estimation. This indicator is shown in the graph (Fig. 4)

Each of the considered methods offers different approaches to accounting for the accumulation of losses, which leads to significantly different final values of the parameter ε under the same training conditions. In the course of such tests, a histogram reflecting the real ratio of error accumulation depending on the chosen composition method was constructed.

The analysis of the obtained results demonstrates a significant scatter of ε values for different composition methods. The rdp method shows the most conservative estimates, reaching a value of 23436.27, reflecting its strict approach to accounting for private losses. In contrast, the gdp method shows the most optimistic estimate with a value of 412.9. The basic and advanced methods occupy an intermediate position with values of 4687.000000 and 8382.1 respectively, while zcdp shows a moderate score of 2672.015.
Developing optimal strategies for choosing a composition method represents a promising area for further research in PPML. Understanding the features and limitations of each method not only allows us to improve current practices, but also to develop more efficient learning algorithms that strike an optimal balance between data privacy and model utility. The most important factor in this area is a very accurate selection of parameters, which often forms the final structure of model building using DP mechanisms.

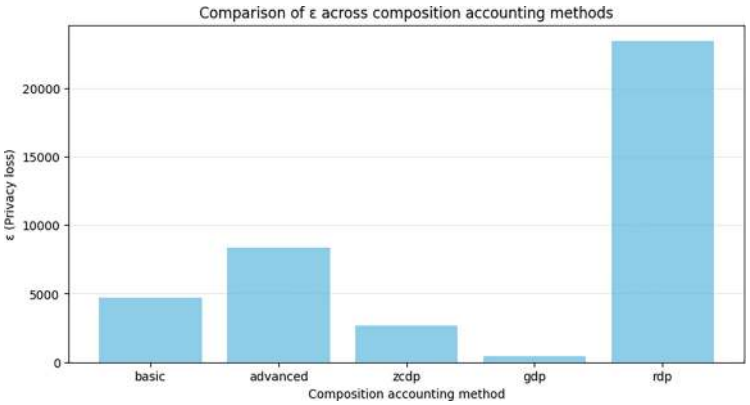


Fig. 4. The degree of error accumulation in different compositional methods.

5 Conclusion

In this study, we examined the concept of differential privacy in the context of iterative machine learning algorithms, with particular emphasis on how different noise-addition mechanisms and privacy composition methods affect the quality and stability of trained models. The theoretical analysis showed that the effectiveness of differential privacy is largely determined by the choice of both the noise mechanism and the composition strategy. The Laplace mechanism offers strong privacy guarantees but is associated with a rapid growth of the leakage parameter ε and a substantial degradation of model accuracy. In contrast, the Gaussian mechanism enables a more flexible trade-off between privacy and utility, especially when combined with modern privacy accounting frameworks such as RDP and zCDP.

The experimental results confirmed that the evolution of privacy loss depends directly on the chosen composition strategy. Methods based on RDP and zCDP demonstrated the best balance between strict privacy guarantees and practical accuracy. More conservative approaches, such as RDP, provide stronger protection, whereas simpler methods yield higher model utility at the cost of a modest increase in cumulative error.

Overall, the study confirms that ensuring differential privacy in machine learning is an inherently multi-parameter problem that requires joint optimisation of the noise mechanism, the composition method, and the training procedure. The results suggest that the most effective practical configuration combines the Gaussian mechanism with RDP-based composition and adaptive optimisers.

Funding The research was supported by the Russian Science Foundation Grant No 25-71-30007, https://rscf.ru/en/project/25-71-30007/.

References

1. Agakov, F., Bonilla, E., Cavazos, J., Franke, B., Fursin, G., O'Boyle, M.F., Thomson, J., Toussaint, M., Williams, C.K.: Using machine learning to focus iterative optimization. In: International Symposium on Code Generation and Optimization (CGO'06), pp. 11–pp. IEEE (2006)
2. Babenko, M., Golimblevskaia, E., Tchernykh, A., Shiriaev, E., Ermakova, T., Pulido-Gaytan, L.B., Valuev, G., Avetisyan, A., Gagloeva, L.A.: A comparative study of secure outsourced matrix multiplication based on homomorphic encryption. Big Data Cogn. Comput. **7**(2), 84 (2023)
3. Bun, M., Steinke, T.: Concentrated differential privacy: simplifications, extensions, and lower bounds. In: Theory of cryptography conference, pp. 635–658. Springer (2016)
4. Dong, J., Roth, A., Su, W.J.: Gaussian differential privacy. J. R. Stat. Soc. Ser. B Stat Methodol. **84**(1), 3–37 (2022)
5. Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming, pp. 1–12. Springer (2006)
6. Hesamifard, E., Takabi, H., Ghasemi, M., Wright, R.N.: Privacy-preserving machine learning as a service. In: Proceedings on Privacy Enhancing Technologies (2018)
7. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
8. Kadam, S.S., Adamuthe, A.C., Patil, A.B.: CNN model for image classification on MNIST and fashion-MNIST dataset. J. Sci. Res. **64**(2), 374–384 (2020)
9. Kairouz, P., Oh, S., Viswanath, P.: The composition theorem for differential privacy. In: International conference on machine learning, pp. 1376–1385. PMLR (2015)
10. Lapina, M., Shiriaev, E., Babenko, M., Istamov, I.: High-speed convolution core architecture for privacy-preserving neural networks. Program. Comput. Softw. **50**(6), 417–424 (2024)
11. LaValley, M.P.: Logistic regression. Circulation **117**(18), 2395–2399 (2008)
12. Mironov, I.: Rényi differential privacy. In: IEEE 30th computer security foundations symposium (CSF), pp. 263–275. IEEE (2017)
13. O'shea, K., Nash, R.: An introduction to convolutional neural networks (2015). arXiv:1511.08458
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32** (2019)
15. Phan, N., Wu, X., Hu, H., Dou, D.: Adaptive laplace mechanism: differential privacy preservation in deep learning. In: 2017 IEEE international conference on data mining (ICDM), pp. 385–394. IEEE (2017)
16. Popescu, M.C., Balas, V.E., Perescu-Popescu, L., Mastorakis, N.: Multilayer perceptron and neural networks. WSEAS Trans. Circuits Syst. **8**(7), 579–588 (2009)
17. Python, W.: Python. Python releases for windows, p. 24 (2021)
18. Shiriaev, E., Nazarov, A., Kucherov, N., Babenko, M.: Analytical review of confidential artificial intelligence: methods and algorithms for deployment in cloud computing. Program. Comput. Softw. **50**(4), 304–314 (2024)
19. Song, S., Chaudhuri, K., Sarwate, A.D.: Stochastic gradient descent with differentially private updates. In: 2013 IEEE global conference on signal and information processing, pp. 245–248. IEEE (2013)

20. Su, X., Yan, X., Tsai, C.L.: Linear regression. Wiley Interdiscip. Rev. Comput. Stat. **4**(3), 275–294 (2012)
21. Vershkov, N.A., Babenko, M.G., Kuchukova, N.N., Kuchukov, V.A., Kucherov, N.N.: Transverse-layer partitioning of artificial neural networks for image classification. Comput. Opt. **48**(2), 312–320 (2024)
22. Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J.: Opacus: user-friendly differential privacy library in pytorch (2021). arXiv:2109.12298

Design of Number-Theoretic Transforms Using Diemitko Numbers

Saveliy Gordeev[1(✉)], Anton Sinitsyn[2], and Maria Lapina[1]

[1] North-Caucasus Federal University, 2 Kulakova Ave, Stavropol 355029, Russia
gordeeff.sava@mail.ru
[2] Russian Academy of Sciences, 109004 Moscow, Russia
antony@email.su

Abstract. This article discusses the implementation of theoretical and numerical transformations on Diemitko numbers, which are special-purpose numbers. Methods for achieving the highest performance when performing transformations using various software and mathematical methods are investigated with the aim of minimising resource consumption. Ready-made tools in the form of libraries and software implemented using the C++ programming language are used as auxiliary means.

Keywords: Homomorphic Encryption · Number Theoretic Transform · Benchmarking · Cryptography · Computer Science · Diemitko numbers

1 Introduction

In recent decades, discrete transformations of numerical sequences have played a key role in digital signal processing, cryptography, and numerical computations. One such transformation that has become widely used is the discrete Fourier transform [18,19], adapted to a finite field (Galois field) [7]. In the context of these computations, various modifications of this transform have been proposed for different types of computations. One effective way to implement this NTT is to use fast transformation algorithms in a system of residue classes modulo a prime number. In this case, modules of a special type that specify the dimension of the field are of particular interest. Among them, the paper [8] considered such as FermatMarcené numbers (GFM) [13], which have a favourable structure for fast calculation of roots of unity. As an alternative, we propose an approach focused on adapting the proposed NTT for *Diemitko numbers* [17], which are a special class of prime numbers of the form $p = qR + 1$ satisfying the Diemitko criterion. These numbers have a number of properties that simplify the construction of roots of unity and ordered cyclic subgroups in the multiplicative group of modulo p subtractions, making them attractive for use in discrete transform algorithms. This case not only speeds up the calculation of forward and inverse transformations, but also ensures the numerical stability of operations in a finite

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 149–161, 2027.
https://doi.org/10.1007/978-3-032-27834-0_15

field, which is especially important when working with large vectors and high levels of accuracy. Thanks to the strictly defined structure of Diemitko numbers and their controlled origin, it is possible to pre-check all necessary arithmetic conditions, which contributes to the reliable implementation of algorithms in applied problems [6]. These properties make this approach promising for use in cryptographic protocols based on fast discrete transformations, for example, in homomorphic encryption schemes [2], as well as in the development of post-quantum cryptography [10–12,20]. The high efficiency of convolution operations and the possibility of reversing the transformation with known parameters allow this method to be integrated into secure systems where both performance and cryptographic strength are strictly required. To achieve the required level of computational cost for the transformations, a number of measures were taken to maximise the results of the NTT, using the data, thereby allowing the results to be aggregated for further research based on the implementation or development of this transformation in existing software tools for high-performance computing in various fields. The C++ programming language was chosen as the main platform for implementing algorithms, as it has a wide range of specialised libraries, low-level memory management, and a high degree of optimisation during compilation. Thanks to its strictly typed structure and support for modular programming [15], C++ allows high execution speeds to be achieved while maintaining the flexibility of the software architecture. Let us consider in more detail the implementation of NTT based on Diemitko numbers. Such numbers have a structure that allows not only to efficiently construct multiplicative cyclic groups, but also to find primitive roots necessary for the implementation of fast transformation in a finite field.

2 Criteria for the Existence of Diémitko Numbers

The construction and generation of these numbers is based on Diemitko's theorem, a mathematical theorem used to test numbers for primality. It is part of the algorithm for generating prime numbers. In particular, Diemitko's theorem establishes the conditions under which a number n can be considered prime. The complete formulation of the theorem is as follows:

Theorem 1. (Diemitko) *Let $n = qR + 1$, where q is a prime number and R is an even number. Then n is prime if the following conditions are met:*

1. *n is not divisible by any prime number less than or equal to $\sqrt{n}$ (or this has been verified separately);*
2. *There exists an integer a such that*

$$a^{qR} \equiv 1 \pmod{n}, \quad \textit{but} \quad a^{R} \not\equiv 1 \pmod{n};$$

3. *$R < 4(q+1)$;*
4. *$n < 4(q+1)^2$.*

If all these conditions are satisfied, then the number n can be considered provably prime. Based on this theorem, an algorithm (Algorithm 1) was proposed for generating the numbers represented. This algorithm demonstrates an effective method for generating numbers with the structure necessary for subsequent use in constructing simple modules. However, it should be noted that its success depends directly on the choice of parameters q and $R_{\max}$. If the range of values of R is not wide enough, the probability of finding a suitable number n is significantly reduced, especially for large values of q.

Algorithm 1 Generating prime numbers using Diemitko's theorem

Input: Prime number q; maximum value $R_{\max}$
Output: Prime number $n = qR+1$ satisfying the conditions of Diemitko's theorem

for $R = 2$ **to** $R_{\max}$ **with step** 2 **do**
  Calculate $n = qR + 1$
  if n is not composite (i.e., has no divisors up to $\sqrt{n}$) **then**
    Choose a random $a \in [2, n-2]$
    if $a^{qR} \bmod n = 1$ **and** $a^{R} \bmod n \neq 1$ **then**
      if $R < 4(q+1)$ **and** $n < 4(q+1)^2$ **then**
        return n
      end if
    end if
  end if
end for
return End of generation

In addition, since verifying the conditions of Dimitkov's theorem involves calculating comparisons modulo a large number, the practical implementation of the algorithm requires optimization in terms of computational complexity. To improve the computational efficiency of the Dimitkov theorem verification implementation, auxiliary procedures that play a key role in optimization were implemented as part of the proposed approach. First, let us consider the process of fast modular exponentiation (Algorithm 2).

Algorithm 2 Fast modular exponentiation

Input: Base $base$, exponent exp, modulus p
Output: Value $base^{exp} \bmod p$

```
result ← 1
base ← base mod p
while exp > 0 do
    if exp mod 2 = 1 then
        result ← (result × base) mod p
    end if
    base ← (base × base) mod p
    exp ← exp/2
end while
return result
```

Furthermore, since verifying the conditions of Diemitko's theorem involves calculating comparisons modulo a large number, the practical implementation of the algorithm requires optimization in terms of computational complexity. To improve computational efficiency, the proposed approach implements an auxiliary procedure (Algorithm 3), which plays a key role in the preliminary filtering of candidates:

Algorithm 3 Naive check for simplicity of a number

Input: Integer n
Output: `true` if n is prime, otherwise `false`

```
if n < 2 then
    return false
end if
if n = 2 or n = 3 then
    return true
end if
if n mod 2 = 0 then
    return false
end if
for i = 3 to ⌊√n⌋ with step 2 do
    if n mod i = 0 then
        return false
    end if
end for
return true
```

The numbers obtained are used to select a simple module for performing the transformation in question, thus ensuring that the conditions necessary for the correct and efficient implementation of NTT in Galois fields are met. Let us

consider in more detail the algorithm for constructing this transformation, identify the main criteria for evaluating the effectiveness of the transformation, and optimize it to achieve the desired result.

3 Theoretical-Numerical Transformation over Diemitko Numbers

As mentioned earlier, the basis for performing NTT over Diemitko numbers is the discrete Fourier transform, which uses the properties of finite fields and the peculiarities of representing prime numbers that satisfy the conditions of Diemitko's theorem. Such numbers guarantee the presence of the necessary structure and order of elements in the field, which ensures the correctness and high efficiency of the transformation calculations.

A feature of applying DFT in this context is the need to choose a suitable prime modulus p that satisfies the strict conditions of Diemitko's decomposition. This ensures the existence of a primitive root of the required order, necessary for constructing the basis of the transform and correctly calculating its coefficients.

For a complete understanding, let us consider the formula for discrete Fourier transform in a finite field:

$$X_k = \sum_{j=0}^{N-1} x_j \cdot \omega^{jk} \pmod{p}, \quad k = 0, 1, \ldots, N-1, \tag{1}$$

where ω is a primitive root of degree N modulo p.
In turn, the formula for the inverse transformation is as follows:

$$X_k = N^{-1} \sum_{j=0}^{N-1} x_j \omega^{-kj} \pmod{p}, \quad k = 0, 1, \ldots, N-1, \tag{2}$$

where ω is a primitive root of degree N modulo p, and N^{-1} is the multiplicative inverse of N modulo p.

As we can see, this formula for constructing NTT is based on the selection criterion for the modulus p, thereby predetermining the structure and properties of the transformation. In the context of our research, the previously considered Diemitko numbers act as a simple module. In addition, subject to the conditions of Diemitko's theorem, it is possible to effectively construct the module p and the corresponding root ω, which significantly reduces the cost of parameter selection and allows us to deterministically ensure the correctness of the transformation. The procedure for selecting ω is a very important element, since, subject to the Diemitko criteria, this number is also subject to typed selection, which in turn imposes a number of necessary conditions for the correct execution of the transformation.

To perform NTT, it is necessary to have a primitive N-th root of unity ω modulo p, i.e., an element that satisfies the following conditions:

$$\omega^N \equiv 1 \pmod{p}, \quad \text{and} \quad \omega^k \not\equiv 1 \pmod{p} \quad \forall\, 0 < k < N. \tag{3}$$

In other words, the order of the element ω modulo p is equal to N if N is the smallest positive integer for which

$$\omega^N \equiv 1 \pmod{p}, \tag{4}$$

and for all lesser degrees

$$\omega^k \not\equiv 1 \pmod{p}, \quad k < N. \tag{5}$$

Since p - prime number, the group Z_p^* has order $p-1$, and the condition $N \mid (p-1)$ is necessary for the existence of such an ω. In the construction based on Diemitko numbers, this condition is automatically satisfied, since by definition $p = qR+1$, where $N \mid (p-1) = qR$.

Thus, the following method is used to find a suitable value for ω:

1. An arbitrary candidate $\alpha \in Z_p^*$ is selected.
2. Check that $\alpha^{(p-1)/N} \not\equiv 1 \pmod{p}$, but $(\alpha^{(p-1)/N})^N \equiv 1 \pmod{p}$.
3. Then $\omega = \alpha^{(p-1)/N}$ and will be the desired primitive root of degree N.

Formally:

$$\omega = \alpha^{(p-1)/N} \pmod{p}, \tag{6}$$

where α is a primitive root modulo p.

When these conditions imposed on NTT are met, the transformation functions correctly, thereby ensuring the desired result in obtaining a powerful mathematical tool suitable for performing calculations in various fields, as well as making it a very promising resource in the field of cryptography development.

However, this method of constructing NTT has a number of disadvantages, which is why it is necessary to optimize it using various existing methods aimed at optimizing computational resources. Among the most effective solutions to this problem is the use of algorithmic techniques similar to the "butterfly" scheme in the Cooley-Tukey FFT algorithm [4], which represents the execution of an elementary step that determines the duration of the Fourier transform calculation, allowing for a significant reduction in the number of multiplication and addition operations when calculating the FFT with Diemitko numbers and its inverse transform. The mathematical implementation of this algorithm is as follows:

$$\begin{aligned} u &= x_{even} + \omega^k \cdot x_{odd} \pmod{p}, \\ v &= x_{even} - \omega^k \cdot x_{odd} \pmod{p}. \end{aligned} \tag{7}$$

where:

– x_{even} and x_{odd}–elements of the input sequence at even and odd positions, respectively;

– ω^k–the corresponding degree of the primitive root ω modulo p;
– u and v–the results of combining elements at the current step of the algorithm.

This step is performed iteratively for each pair of elements at each level of division, which together forms a complete scheme for fast calculation of NTT. As part of TCP calculations, this procedure is applied iteratively, dividing the input sequence into blocks, which reduces the overall computational complexity from $O(N^2)$ to $O(N \log N)$.

However, in practice, it is necessary to verify the correctness of the algorithm and ensure that the original data can be recovered from the output values before applying the transformation. In practice, difficulties may arise due to the limitations imposed by the finite field, in particular, in terms of number handling and working with modulo division.

4 Experimental Research

This study aims to optimize theoretical-numerical conversion over Diemitko numbers using various software tools and methods. The main goal of the experiments is to verify the correctness of the implemented algorithms, evaluate their performance, and analyze the impact of the parameters N, p, and ω on the efficiency and accuracy of calculations. A scenario was proposed in which an approach based on the automatic selection of the dimension of the modulus p depending on the length of the input vector and the corresponding primitive root is implemented. This transformation was then modified and compared with the software implementation of NTT offered by libraries in the C++ programming language. In the course of this, the most optimal approach was identified, combining several methods and optimized to achieve the fastest possible transformation.

To verify the dependence of the conversion time on the length of the input vector, an experimental dependence was compiled, during which the corresponding values of the time component of the conversion were identified using dynamic field selection. The result is shown in the graph (Fig. 1).
In turn, the dependency graph (Fig. 2) obtained as a result of backpropagation looks as follows:
The graph shows how the time of the transformation increases with the growth of the input vector dimension. However, we can see that the execution time of the transformation based on Diemitko numbers is several times faster than the transformation using generalized Fermat-Mersen numbers. This result shows that using Diemitko numbers as a simple modulus is a very promising solution for minimizing the conversion time. To improve the efficiency of theoretical-numerical conversion based on Diemitko numbers, several key optimization methods aimed at reducing execution time and improving computational stability were considered and implemented as part of the study:

1. **Using bitwise operations**. [1] In modulo operations, as well as in the implementation of fast exponentiation loops, significant acceleration is achieved

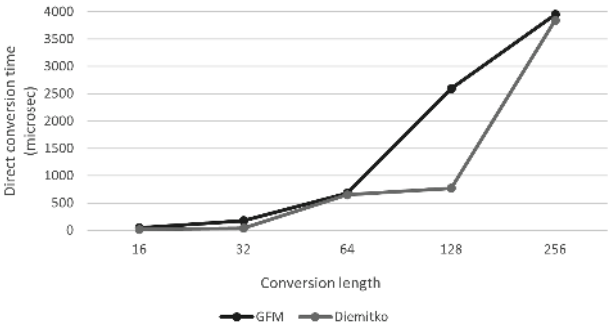


Fig. 1. Time dependence of direct conversion on the length of the input vector

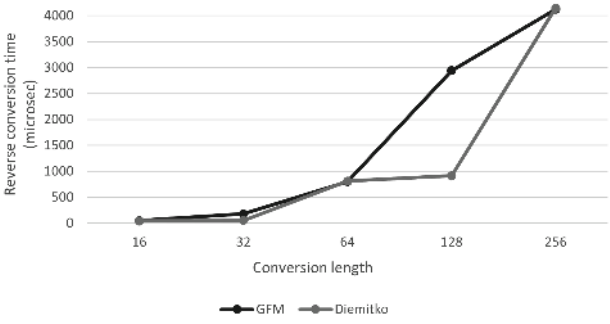


Fig. 2. Time dependence of the inverse transformation on the length of the input vector

through the use of bit shifts and logical operations. For example, division by two in loop iterations is implemented by a right shift (`exp >> 1`), which is significantly faster than classic division. Such optimizations are critical when working with large amounts of data and allow you to speed up the most "bottlenecked" parts of the conversion implementation.

2. **Application of the CooleyTukey algorithm. Butterfly algorithm** To perform the transformation, a scheme of dividing the input vector into subblocks followed by recalculation of the coefficients, known as the "butterfly algorithm," is effectively applied. This allows the task to be recursively broken down into smaller subtasks and the transformation to be performed with a computational complexity of $O(N \log N)$ instead of the naive $O(N^2)$. Mathematically, this approach is based on factorizing the size of the transformation and applying elementary operations with pairs of elements at each level of recursion in stages.
3. **Division optimization using Barrett's algorithm**. [9] When performing modulo operations, especially in cyclic multiplications, the standard division operation causes significant overhead. Barrett's algorithm avoids full division by precomputing an approximate inverse of the modulus and then using mul-

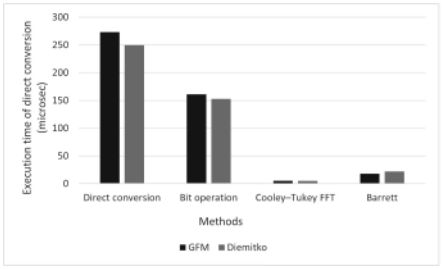
tiplication and shifts to quickly obtain the remainder. This is particularly effective in cases where the modulus p remains constant for a large number of operations.

Also, as mentioned earlier, a number of specialized libraries were considered during the current stage, which, according to theoretical and practical estimates, can significantly reduce the time required for transformations and simplify the implementation of low-level operations:

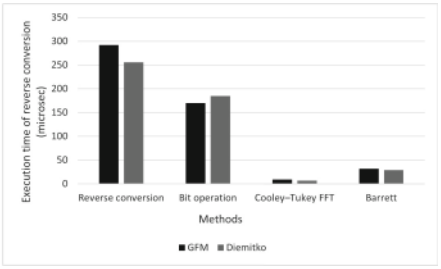
1. **HEXL (High-Performance Encryption Library)** [5]. Developed by Intel, this library is optimized for architectures supporting AVX-512 and is designed to perform numerical operations in cryptographic algorithms, including fast Fourier transforms in finite fields. HEXL provides high-performance implementations of modular multiplication operations, including the Barrett algorithm, and makes efficient use of the processor's SIMD instructions [22]. The library is actively used in homomorphic encryption, but can also be adapted for TCP tasks.
2. **Boost.Multiprecision** [14]. A library from the Boost C++ collection that provides support for integers and floating-point numbers of arbitrary precision. It is especially useful when you need to perform calculations with numbers that exceed standard machine precision. Thanks to its abstract interface, you can easily switch between different implementations, which is convenient for analyzing performance and accuracy under different conditions.
3. **NTL (Number Theory Library)** [3]. A classic library widely used in scientific and research tasks related to number theory. NTL provides a rich set of functions for working with finite fields, matrices, polynomials, and transforms. It is a reliable basis for implementing number theory transformations, as well as for generating prime numbers, including using Diemitko-type criteria. NTL can be integrated with GFM to improve accuracy and performance.

Comparative experiments were conducted to evaluate the effectiveness of the described approaches and software implementations. Taking into account the specifics of the changes, statistics were collected as a result of performing the transformation with modifications. The execution time of the transformations was measured depending on the length of the input vector and the optimization technique used. This resulted in a comparative graph (Fig. 3) of execution times for different methods.

However, in addition to the modifications we have proposed, it is also worth conducting a comparative analysis of existing software methods in the form of libraries for the C++ programming language. This resulted in a graph comparing the methods offered by the libraries (Fig. 4). The combined use of these methods made it possible to significantly reduce the conversion time, increase the efficiency of computing resources, and ensure the scalability of the solution as the input data size increased. Thus, during the analysis of conversion implementation methods, a benchmark option was identified that combines high performance, stability, and ease of integration into computing systems. This result



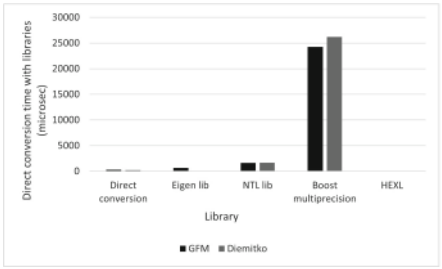
(a) Time dependence of direct conversion on input vector length, taking into account modifications



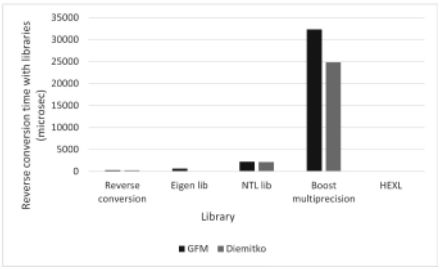
(b) Time dependence of the inverse transformation on the length of the input vector, taking into account modifications

Fig. 3. Transformation execution time, taking into account modifications: **a** Forward transformation, **b** Reverse transformation

was achieved by applying optimizations at the bit operation level and replacing expensive division operations with fast analogues. A "butterfly" scheme was also implemented in the Cooley-Tukey algorithm to effectively divide the task and reduce the number of operations.



(a) Direct conversion time taking into account library tools



(b) Reverse conversion time considering library tools

Fig. 4. Transformation execution time with modifications proposed by libraries: **a** Forward transformation, **b** Reverse transformation

A separate branch of research is the optimization of methods offered by the specialized Intel HEXL library, which, as experimental measurements have shown, provides one of the lowest results. A separate branch of research is the optimization of methods offered by the specialized Intel HEXL library, which, as experimental measurements have shown, provides one of the lowest conversion execution times due to deep optimization for modern architectures and SIMD instructions.

5 Conclusion

The results of this research open up prospects for the use of Diemitko number transformations in many applied problems. One of the key areas of application is cryptography and homomorphic encryption, where transformations over finite fields, especially with special types of modules, can significantly reduce computational costs. Thanks to the high performance and optimized structure of Diemitko numbers, such transformations are particularly attractive for building cryptographic protocols that require high data processing speeds.

Another area where theoretical-numerical transformation with such modules may be in demand is digital signal processing. Numerical conversion based on Diemitko numbers is used in signal compression, restoration, and filtering [16], especially in cases where working with real numbers is impossible or undesirable. The use of Diemitko numbers allows for the implementation of efficient algorithms in conditions of limited computational resources or strict time requirements.

In addition, the prospects for using the results obtained may be related to high-performance computing, including simulation, numerical methods in scientific computing, and machine learning [21]. The availability of a fast and scalable transformation can be integrated into linear algebraic computations, polynomial operations, pseudo-random number generation, and other tasks where fast operation with finite fields is important.

Ultimately, the study confirms both the theoretical validity of using Diemitko numbers in NTT and their practical applicability in modern computing systems. The implementation of the described optimizations not only makes this transformation competitive with existing approaches, but also opens up opportunities for further development and adaptation in related disciplines.

Funding: The research was supported by the Russian Science Foundation Grant No 25-71-30007, https://rscf.ru/en/project/25-71-30007.

References

1. Aaronson, S., Grier, D., Schaeffer, L.: The classification of reversible bit operations (2015). https://doi.org/10.48550/arXiv.1504.05155, arXiv:1504.05155 [quant-ph]
2. Acar, A., Aksu, H., Uluagac, A.S., Conti, M.: A survey on homomorphic encryption schemes: theory and implementation. ACM Comput. Surv. **51**(4), 1–35 (2019). https://doi.org/10.1145/3214303, https://dl.acm.org/doi/10.1145/3214303
3. Aguilar-Melchor, C., Barrier, J., Guelton, S., Guinet, A., Killijian, M.O., Lepoint, T.: NFLlib: NTT-based fast lattice library. In: Sako, K. (ed.) Topics in Cryptology-CT-RSA 2016, vol. 9610, pp. 341–356. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29485-8_20, http://link.springer.com/10.1007/978-3-319-29485-8_20
4. Bekele, A.J.A.A.: Cooley-Tukey FFT Algorithms. Advanced Algorithms (2016)
5. Boemer, F., Kim, S., Seifu, G., De Souza, F.D.M., Gopal, V.: Intel HEXL: Accelerating homomorphic encryption with intel AVX512-IFMA52. In: Proceedings of the 9th Workshop on Encrypted Computing & Applied Homomorphic Cryptography, pp. 57–62. ACM, Virtual Event Republic of Korea (2021). https://doi.org/10.1145/3474366.3486926, https://dl.acm.org/doi/10.1145/3474366.3486926

6. Boussakta, S., Holt, A.G.J.: Number theoretic transforms and their applications in image processing. In: Advances in Imaging and Electron Physics, vol. 111, pp. 1–90. Elsevier (1999). https://doi.org/10.1016/S1076-5670(08)70216-7, https://linkinghub.elsevier.com/retrieve/pii/S1076567008702167
7. Dickson, L.E.: Linear Groups with an Exposition of the Galois Field Theory. Cosimo, New York (2007)
8. Dimitrov, V.S., Cooklev, T.V., Donevsky, B.D.: Generalized Fermat-Mersenne number theoretic transform. IEEE Trans. Circuits Syst. II **41**(2), 133–139 (1994). https://doi.org/10.1109/82.281844, http://ieeexplore.ieee.org/document/281844/
9. Garg, H.K., Xiao, H.: New residue arithmetic based Barrett algorithms: modular integer computations. IEEE Access **4**, 4882–4890 (2016). https://doi.org/10.1109/ACCESS.2016.2599269, http://ieeexplore.ieee.org/document/7539558/
10. Gladkov, A., Kuchukov, V., Babenko, M., Tchernykh, A., Berezhnoy, V., Drozdov, A.Y.: Modified error detection and localization in the residue number system. Program. Comput. Softw. **48**(8), 598–605 (2022). https://doi.org/10.1134/S0361768822080126, https://link.springer.com/10.1134/S0361768822080126
11. Kumar, M., Pattnaik, P.: Post quantum cryptography(PQC)-an overview. In: 2020 IEEE High Performance Extreme Computing Conference (HPEC), pp. 1–9. IEEE, Waltham, MA, USA (2020). https://doi.org/10.1109/HPEC43674.2020.9286147, https://ieeexplore.ieee.org/document/9286147/
12. Lapina, M.A., Shiriaev, E.M., Babenko, M.G., Istamov, I.: High-speed convolution core architecture for privacy-preserving neural networks. Program. Comput. Softw. **50**(6), 417–424 (2024). https://doi.org/10.1134/S0361768824700282, https://link.springer.com/10.1134/S0361768824700282
13. Lee, S.C., Lu, H.: Fast convolution using generalized Fermat/Mersenne number transforms. In: ICASSP-88. International Conference on Acoustics, Speech, and Signal Processing, pp. 1910–1913. IEEE, New York, NY, USA (1988). https://doi.org/10.1109/ICASSP.1988.197000, http://ieeexplore.ieee.org/document/197000/
14. Maddock, J., Kormanyos, C., et al.: Boost multiprecision (2018). http://lars.mec.ua.pt/public/LAR_Projects/Perception/2013_PedroSilva/Dissertacao/Code_bank/boost_1_53_pdf/multiprecision.pdf
15. Madison, S., Gifford, J.: Modular programming: novice misconceptions. J. Res. Technol. Educ. **34**(3), 217–229 (2002)
16. Najim, M.: Modeling, Estimation and Optimal Filtering in Signal Processing. John Wiley & Sons, London (2008)
17. Peng, L., Hu, L., Lu, Y., Wei, H.: An improved analysis on three variants of the RSA cryptosystem. In: Chen, K., Lin, D., Yung, M. (eds.) Information Security and Cryptology, vol. 10143, pp. 140–149. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54705-3_9, http://link.springer.com/10.1007/978-3-319-54705-3_9
18. Sundararajan, D.: The Discrete Fourier Transform: Theory, Algorithms and Applications. World Scientific Pub. Co, Singapore (2001)
19. Valuev, G., Valueva, M., Babenko, M., Tchernykh, A.: Digital filter architecture based on modified winograd method F(2x2, 5x5) and residue number system. IEEE Access **11**, 26807–26819 (2023). https://doi.org/10.1109/ACCESS.2023.3258689, https://ieeexplore.ieee.org/document/10075443/
20. Vershkov, N., Babenko, M., Tchernykh, A., Kuchukov, V., Kucherov, N., Kuchukova, N., Drozdov, A.Y.: Optimization of artificial neural networks using wavelet transforms. Program. Comput. Softw. **48**(6), 376–384 (2022). https://doi.org/10.1134/S036176882206007X, https://link.springer.com/10.1134/S036176882206007X

21. Xu, R., Baracaldo, N., Joshi, J.: Privacy-preserving machine learning: methods, challenges and directions (2021). https://doi.org/10.48550/ARXIV.2108.04417, https://arxiv.org/abs/2108.04417
22. Zhou, J., Ross, K.A.: Implementing database operations using SIMD instructions. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, pp. 145–156. ACM, Madison, Wisconsin (2002). https://doi.org/10.1145/564691.564709, https://dl.acm.org/doi/10.1145/564691.564709

Application of Artificial Intelligence in Telecommunications

Svetlana Govorova[1](✉), Egor Govorov[1], Malvina Karabasheva[2], and Neelakandan S[3]

[1] North Caucasus Federal University, Stavropol, Russia
mitnik2@yandex.ru
[2] Financial University under the Government of the Russian Federation, Moscow, Russia
[3] Department of Computer Science and Engineering, R.M.K Engineering College, Chennai, India

Abstract. The article is devoted to the application of artificial intelligence (AI) and machine learning (MO) technologies to solve key tasks in modern telecommunication networks. The classification of tasks such as analysis, management and traffic generation is considered, and the need for the introduction of intelligent systems to overcome the limitations of traditional approaches is substantiated, especially in the context of increasing data volume and complexity of network infrastructures. Practical examples demonstrate the effectiveness of various MO methods: recurrent neural networks (NARX) for predicting delays, multiple regression models (linear, power-law, SVR, etc.) for predicting queue size in buffers, and ensemble algorithms (Random Forest, XGBoost) for detecting network anomalies. The conducted research shows that the use of AI not only increases the accuracy and efficiency of network management, but also allows you to automate routine operations, reducing the burden on the operator and minimizing the risk of human error. It is concluded that AI serves as a powerful tool for optimizing resources, ensuring quality of service (QoS), and improving the overall stability of telecommunications systems.

Keywords: Artificial intelligence · Communication networks · Traffic management · Machine and deep learning · Neural network · RMSE · MAPE · MAE

1 Introduction

The development of communication networks in the first and second decades of the 21st century takes place against the background of global changes in telecommunications technologies, which have a non-trivial impact on both the architecture of communication networks and the main indicators of the functioning of communication networks. First of all, these are ultra-dense and ultra-low latency communication networks [1]. The emergence of the former is due to the development of the concept of the Internet of Things, while the latter are a response to the challenges posed by the concept of the Tactile Internet [1]. The dominant concept of the Internet of Things has acted as a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 162–170, 2027.
https://doi.org/10.1007/978-3-032-27834-0_16

catalyst for the development of two key technologies: fifth-generation communication networks and ultra-dense networks. The operation of modern communication networks places increased demands on the quality of network service, which necessitates finding a balance between the growing volume of traffic, quality of service (QoS) requirements and available network resources [2, 3]. Existing approaches are not flexible and operational enough to adapt to dynamic and non-linear changes in the network architecture. In this regard, there is an urgent need to implement systems based on artificial intelligence (AI) and machine learning [4]. They can perform tasks like humans, but without the inherent disadvantages of fatigue and inattention. It is important to understand that AI does not displace a specialist, but serves him as a tool that takes on routine and complex calculations in order to increase the efficiency of human work. Artificial intelligence technologies can be used to solve a wide range of tasks in the telecommunications industry, including traffic analysis and management [5].

2 Tasks of Analysis and Traffic Management of Telecommunication Networks

In modern telecommunications, there is a steady and rapid growth in the volume of transmitted traffic, which is due to a complex of interrelated factors. The main driver was the widespread distribution of video content: streaming services in 4 K/8 K format, videoconferencing and platforms with short videos create a constant and significant load on networks. At the same time, the mass adoption of the Internet of Things (IoT) – from smart homes and industrial sensors to telematics systems – provides a background but large-scale data flow from millions of connected devices [5].

The development of wireless technologies, in particular the deployment of fifth-generation (5G) networks, acts not only as a response to growing demand, but also as a catalyst for new consumption [6]. High speeds and low delays encourage the use of mobile Internet as the main one, opening the way for demanding services such as cloud gaming and interactive online platforms. The Ericson Mobility Report indicates that mobile data traffic will increase threefold between 2025 and 2029, creating a need for network management automation, which will lead to an increase in the use of artificial intelligence automation technologies [1].

The main task of the communication network is the provision of services, that is, traffic maintenance. Network performance is assessed using quantitative quality of service (QoS) indicators, which must meet regulatory criteria. These criteria may vary depending on the type of communication service.

When building and operating communication networks, the task of choosing a compromise between three main groups of parameters is solved: traffic, quality of service and volume of resources. The tasks related to traffic analysis and management are classified into three categories: analysis, management and traffic generation [7].

The traffic analysis process covers continuous monitoring, data classification, and anomaly detection. It also aims to ensure the quality of service by monitoring metrics such as latency, jitter, and packet loss. In addition, its tasks include evaluating various traffic parameters and investigating its frequency and long-term dynamics. The solution

of traffic analysis tasks includes several stages: the stage of statistical data collection; preliminary data processing; statistical analysis; application of the results obtained [8].

At the stage of preliminary processing of statistical data, it is possible to use AI systems trained during the operation of the monitoring system.

During statistical traffic analysis, network traffic data is examined and evaluated in order to identify patterns, trends, and anomalies. Traffic intensity, packet delays and losses, and time patterns are analyzed. Using well-known mathematical tools does not always provide the right solutions for such problems. The use of AI technologies makes it possible to solve these problems more efficiently.

The tasks of traffic management consist in actively influencing traffic to achieve specific goals: prioritization, load balancing, traffic shaping, blocking unwanted traffic, congestion control and traffic modification [7]. Packet loss is often a natural result of network congestion. It occurs when the total volume of traffic that all network participants are trying to transmit exceeds its physical or logical bandwidth in some area. As long as the traffic is not too intense, the buffers successfully smooth out short-term spikes, and the user does not notice problems. However, if the overload continues, a key event occurs — buffer overflow. Traffic restriction methods are used to deal with congestion. Such traffic restriction methods prevent the network from reaching the point of congestion and buffer overflow. When using such traffic restriction methods, the choice of dropped packets poses a certain difficulty. Artificial intelligence is fundamentally changing the paradigm of network traffic management, shifting it from static rules to dynamic, self-learning adaptation.

The task of traffic generation is to create artificial, test traffic that simulates real traffic and can be used for stress testing and throughput verification, stability testing, QoS policy validation, performance evaluation of new equipment or software [7]. Using probability theory and queuing theory, it is quite difficult to take into account the main characteristics of traffic, such as the distribution of intervals between packets, packet sizes, batch flow and the property of self-similarity [2]. Artificial intelligence allows you to generate synthetic network traffic that reliably reproduces all the key properties of real data.

3 Tasks Solved by AI Systems and their Classification

The tasks solved using AI systems in communication networks can be conditionally classified as follows [9].

Classification tasks are the most sought–after class of tasks when using AI. The solution to the problem is to get an answer about belonging to a certain class based on a set of characteristics (in AI technologies it is described as learning with a teacher).

Regression tasks are forecasting tasks based on a sample of some source (most often statistical) data. The solution is a real number representing the predicted value of some feature.

Clustering tasks are the distribution of data into groups based on certain criteria (described in AI technologies as unsupervised learning).

The tasks of dimensionality reduction are to reduce the initial number of features to a smaller number.

The objectives of outlier detection are to isolate data that differs from normal cases. The task is similar to the classification task, the difference lies in the small amount of training data containing outliers (abnormal values).

4 Using Machine and Deep Learning to Predict Latency and Throughput

In [10], to solve the problem of delay prediction in fifth-generation communication networks, it was proposed to use a nonlinear autoregressive recurrent neural network with external NARX inputs designed for time series prediction, and three learning algorithms (trainlm, trainrp, traincgf), which are variations of the Backpropagation method. The effectiveness of these learning algorithms is compared.

For comparison, two metrics for evaluating the quality of models were used: RMSE (Root Mean Square Error)–the root of the root mean square error, and MAPE (Mean Absolute Percentage Error)–the average absolute percentage error [11].

RMSE is one of the most common metrics for evaluating regression models. It is calculated using the formula:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum (y_i - \hat{y}_i)^2}$$

MAPE expresses the error as a percentage of the actual values and is calculated using the formula:

$$MAPE = \frac{1}{n} \cdot \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%$$

where

n is the number of observations,

y_i is the actual value,

y_i is the predicted value.

Using both metrics allows you to get a more complete picture and help you evaluate the model from different angles: RMSE shows the absolute magnitude of the error, and MAPE shows the relative magnitude.

The simulation results are presented in Table 1. In this case, both one-step forward and several-step forward forecasting was used.

Based on the data presented in the table, it can be concluded that nonlinear recurrent NARX networks and the Levenberg-Marquardt learning algorithm with backpropagation have high accuracy in predicting delays.

5 Using Machine Learning Methods to Predict Queue Size

One of the problems of telecommunication networks is the inefficient use of node and channel resources when servicing self-similar traffic. The main reason leading to buffer overflow is the presence of long-term dependence in network traffic due to its self-similarity.

Table 1. The effectiveness of learning algorithms in predicting traffic transmission delays.

Learning algorithm	Forecasting one step ahead		Forecasting a few steps ahead	
	RMSE	MAPE	RMSE	MAPE
Trainlm (Levenberg-Marquardt algorithm)	0.0551	0.0429	0.0521	0.1301
Traincgf (Fletcher-reeves conjugate gradient algorithm)	0.2163	0.5178	0.3704	3.2367
Trainrp	1.4734	0.7245	0.6996	4.7953

Regression analysis is an important statistical modeling tool that is widely used in the telecommunications industry to solve a variety of practical problems. This method allows you to establish quantitative relationships between key network parameters and build accurate forecasts of infrastructure development based on these relationships [12]. One of the main uses of regression analysis has become network load forecasting, where models are built based on historical traffic data that can predict the volume of transmitted data, taking into account seasonal fluctuations, time of day, and the appearance of new services. This allows operators to plan infrastructure upgrades in advance and optimally allocate resources.

In [13], using machine learning methods, models were developed to predict queue size depending on the Hurst index. To build a dependence of the size of the buffer memory queue of switching nodes on the Hurst exponent, the following models and machine learning algorithms were used to solve regression and classification problems: linear model, linear model in a reclining space, models based on hyperbolic regression, power regression, exponential regression, isotonic regression, and the support vector method [14].

The following quality metrics were used: the coefficient of determination R2, the RMS regression error RMSE, the average absolute error MAE, and the estimated loss value.

Then, a comparative analysis of the models was carried out (Table 2) [13] and the best method was determined for predicting the queue size based on the value of the Hurst index for ways to convert input and output traffic with equal medians and mathematical expectation.

The analysis showed that the best predictive ability for the method of converting the input and output packet streams with equal mathematical expectation is a model constructed using the method of support vectors.

For the method of converting input and output traffic when the medians are equal, the best results are obtained using power-law regression.

Table 2. Comparative characteristics of regression methods.

Regression methods	Estimated loss value		Efficiency	
	Mathem. expectation	Median	Mathem. expectation	Median
Linear regression	48.404104	4.976265	0,952730	0,995140
Hyperbolic regression 1	55.25577	4.649444	0,946039	0,995459
Hyperbolic regression 2	28.189645	6.141958	0,972471	0,994001
Power regression	51.394345	5.418426	0,949810	0,994708
Exponential regression	45.25324	5.5233289	0,95580	0,994606
Isotonic regression	19.75032	4.5328338	0,980712	0,995573
SVR	15.43061	5.5961137	0,984931	0,994535
SVR on a logarithmic scale	19.950616	3.580060	0,980516	0,996503

6 Using Machine Learning Techniques for Machine Learning Models to Detect Network Anomalies

Anomalies in Internet traffic are unpredictable, unusual, or deviating from established norms of online interaction. They may indicate the presence of malware, hacks, or other unwanted events. Machine learning (ML) makes it possible to detect anomalies in network traffic, which is important for ensuring cybersecurity [15].

It is possible to detect network anomalies by using various anomaly detection methods: behavioral methods, computational intelligence methods, and machine learning methods [16].

In [17], two categories of machine learning models are compared to identify network anomalies: linear models (logistic regression) and ensemble methods, including running implemented by the random forest algorithm and boosting (XGB classifier, XGB regressor) [18].

Table 3 shows the results of a study of machine learning models with randomly selected initial parameters.

Table 3. Results of the study of machine learning models.

The machine learning algorithm	Accuracy of predictions (AUC ROC)	Accuracy of predictions (cross-validation)	Learning rate	Speed of prediction
Logistic regression	83,21%	73,56%	5,07 s	28 мкс
Random Forest	97,74%	96,41%	2 min 36 s	2,34 с

(*continued*)

Table 3. (*continued*)

The machine learning algorithm	Accuracy of predictions (AUC ROC)	Accuracy of predictions (cross-validation)	Learning rate	Speed of prediction
XGB regressor	84,13%	80,34%	1 min 24 s	1,31 c
XGB classifier	73,49%	67.82%	15 s	1,44 c

7 Conclusion

There are various criteria for evaluating the effectiveness of machine learning models in solving the problem of network traffic profiling. In accordance with the specifics of the task, the most appropriate one was chosen: AUC ROC – the area under the error curve.

According to the results presented in Table 3, the model implemented using the Random Forest algorithm turned out to be of the highest quality (97.74% in accordance with the ROC curve, 96.41% in accordance with the cross-validation check).

Conclusion. The analysis confirms the high efficiency and promising application of artificial intelligence and machine learning technologies in the telecommunications industry. Solving the problems of traffic analysis, management and generation using MO methods allows us to achieve significant progress in improving the quality of service, optimal use of resources and ensuring network security.

The key results of the work demonstrate that:

1. To predict network parameters such as latency, the most accurate results are shown by complex models similar to the recurrent neural network NARX with the Levenberg-Marquardt algorithm.
2. In queue size prediction tasks, the best predictive ability is achieved using the support vector machine (SVR) and power regression, depending on traffic conversion conditions.
3. To detect network anomalies, ensemble methods, in particular the Random Forest algorithm, demonstrate excellent accuracy (up to 97.74%) compared to linear models.

Thus, the integration of AI systems into the operation of communication networks is not just a trend, but an objective necessity for building flexible, efficient and sustainable telecommunications infrastructures of the future. Further development is associated with the creation of more complex autonomous systems capable of continuous learning and adaptation in real time, which will fully unlock the potential of 5G/6G technologies and the Internet of Things.

References

1. Ericsson Mobility Report: (2024). https://ict.moscow/analytics/ericsson-mobility-report-november-2024/?ysclid=mhkxufnma245659297
2. Karmeshu, Sharma, S.: Long tail behavior of queue lengths in broadband networks: Tsallis entropy framework. Netw. Internet Arch. (cs.NI). https://doi.org/10.48550/arXiv.1012.2464

3. Karakus, M., Durresi, A.: Quality of service (QoS) in software defined networking (SDN): a survey. J. Netw. Comput. Appl. **80**, 200–218 (2017). https://doi.org/10.1016/j.jnca.2016.12.019
4. Bakker, J.N., Ng, B., Seah, W.K., Traffic, P.A.: Classification with machine learning in a live network. In: Proceedings of IFIP/IEEE Symposium on Integrated Network and Service Management (IM), pp. 488–493 (2019)
5. Анализ и прогнозирование трафика современных телекоммуникационных систем на основе методов искусственного интеллекта / Д. В. Кутузов, А. В. Осовский, Д. В. Старов [и др.]. https://doi.org/10.24143/2072-9502-2024-1-73-87
6. Latah, M., Toker, L.: Artificial intelligence enabled software defined networking: a comprehensive overview. IET Netw. **8** (2018) https://doi.org/10.1049/iet-net.2018.5082
7. Get'man, A.I., Markin, Y.V., Obydenkov, D.O., Padaryan, V.A., Tikhonov, A.Y.: Methods of presenting the results of network traffic analysis. Tr. ISP RAN/Proc. ISP RAS. **28**(6), 103–110 (2016). https://doi.org/10.15514/ISPRAS-2016-28(6)-7
8. Alcock, S., Nelson, R.: Libprotoident: Traffic Classification Using Lightweight Packet Inspection, Technical report. University of Waikato (2013) https://www.researchgate.net/publication/268404135_Libprotoident_Traffic_Classification_Using_Lightweight_Packet_Inspection#fullTextFileContent. Accessed 01 Nov 2025
9. Abdellah, A.R., Alshahrani, A., et al.: Performance estimation in V2X networks using deep learning-based m-estimator loss functions in the presence of outliers. Symmetry. **13**(11) (2021). https://doi.org/10.3390/SYM13112207
10. Бородин, А. С. Искусственный интеллект в сетях связи пятого и последующих поколений / А. С. Бородин, А. Н. Волков, А. С.А. Мутханна, А. Е. Кучерявый. // Э лектросвязь: ежемесячный научно-технический журнал по проводной и радиосвязи, телевидению, радиовещанию. 2021. N 1. С. 17–22.
11. Управление трафиком телекоммуникационных систем с использованием прогностических методов искусственного интеллекта. Д.В. Кутузов, А.В. Осовский, Д.В. Старов, К.В. Перова Инженерный вестник Дона, №1 (2024). URL: http://ivdon.ru/ru/magazine/archive/n1y2024/8920
12. Advances in Machine Learning/Deep Learning-Based Technologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis. Tsihrintzis G. A.,Virvou M., Jain L. C., Cham: Springer International Publishing. 231 (2022)
13. Linets, G.I., Voronkin, R.A., Govorova, S.V.: Evaluation of the efficiency of reference machine learning models for buffer memory prediction when transforming a self-similar input stream of packets into a stream having exponential distribution under the condition of equality of mathematical expectations and median flows. Mod. Sci. Innovations. **3**, 20–32 (2022) https://doi.org/10.37493/2307-910X.2022.3.2
14. Boutaba, R. et al.: A comprehensive survey on machine learning for networking: evolution applications and research opportunities. J. Internet Serv. Applications. **9**, 1–99 (2018). https://doi.org/10.1186/s13174-018-0087-2
15. Adebayo, A.S., Chukwurah, N., Ajayi, O.O.: Artificial intelligence and machine learning algorithms for advanced threat detection and cybersecurity risk mitigation strategies. Eng. Technol. J. **10**(3), 4080–4094 (2025) https://doi.org/10.47191/etj/v10i03.18
16. Apruzzese, G., Andreolini, M., Ferretti, L.: Modeling Realistic Adversarial Attacks against Network Intrusion Detection Systems (2022). https://doi.org/10.1145/3469659

17. Comparative analysis and development of recommendations for the use of machine learning methods to identify network traffic anomalies in the development of a subsystem for user behavioral analysis. AISMA-2023: International Workshop on Advanced Information Security Management and Applications (AISMA 2023), 17 November 2024, Pp 74–84. Lecture Notes in Networks and Systems–2024–1207 LNNS, pp. 74–84. 10.1007/978-3-031-77229-0_8 Svetlana Govorova, Egor Govorov,Vitalii Lapin, Mary Anita
18. Das, S., Nene, M.J.: A survey on types of machine learning techniques in intrusion prevention systems. (2017). https://doi.org/10.1109/WiSPNET.2017.8300169

Neural Network Method for Solving the System of Differential Equations of the Dynamic Input-Output Balance

Natalia Gudieva(✉)

North Caucasus Center for Mathematical Research, North Caucasus Federal University, Stavropol, Russia
gudieva82@bk.ru

Abstract. An input-output model is an economic tool that enables detailed analysis and forecasting of economic activity in a country or region. From a practical standpoint, it is a system of differential equations that describes the interrelationships among various sectors of the economy. The effectiveness of economic planning and management depends on the accuracy of its solution. Currently, analytical solutions enable the construction of sufficiently accurate short-and medium-term forecasts. However, given the scale of the problem being addressed, even a slight inaccuracy can result in significant financial losses. This paper formulates an optimization problem and proposes a neural network method for refining analytical solutions. Its effectiveness is demonstrated with a specific example: the process reduced the mean relative percentage error from 16.72% to 0.001% without engaging substantial computational resources and within an acceptable time frame.

Keywords: Dynamic input-output model · Singular value decomposition · Matrix exponential · Artificial neural network

1 Introduction

One approach to studying dynamic economic systems is to represent the model within an input–output framework using a system of differential Equations [1, 2]. The input–output model is an analytical tool that enables detailed examination and forecasting of economic activity in a country or region [3]. In practice, it can be formulated as a system of differential equations that describes intersectoral linkages across the economy.

The central purpose of an input–output model is to identify which goods and services are produced by each industry, which resources are required for that production, and which goods and services are used as intermediate inputs by other industries [4, 5]. By quantifying these flows, the model provides valuable information about the interconnections and dependencies among economic sectors [6]. It helps determine the specific combinations of resources needed to produce different categories of output, for example, how much energy, raw materials, labor, and other inputs are required to generate a given final product.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 171–181, 2027.
https://doi.org/10.1007/978-3-032-27834-0_17

Comparing input-output tables across years enables tracking changes in resource allocation over time. This is particularly useful for studying long-run trends and structural shifts in the economy. For instance, such comparisons can reveal how energy intensity has evolved, or how labor requirements in a particular industry have changed.

Without an input–output framework, analyzing structural change becomes substantially more difficult. The model provides a comprehensive representation of how resources and production volumes are distributed across sectors. This type of analysis is essential for policy and managerial decision-making, for example, when identifying priority industries for development or planning the efficient use of scarce resources.

Information derived from input–output analysis is therefore essential for understanding economic mechanisms and anticipating the consequences of shocks. It supports the evaluation of how changes in one sector propagate to others and helps identify industries pivotal to economic growth. In this way, input–output analysis can improve planning, optimize resource allocation, and support more informed financial decisions.

Several approaches are used to construct input–output models, including the technical-coefficient (input–output coefficient) method, the production-function approach, economic modeling techniques, and expert-judgment methods [7]. Input–output balances can also be employed to forecast economic expansion or contraction [8–10]. For example, if one industry grows more rapidly than others, it can increase demand for upstream and downstream products [11]. If related industries cannot expand output sufficiently to meet this demand, supply constraints may emerge, creating broader economic difficulties [8, 12].

Overall, the input–output model is a key instrument for economic analysis and forecasting, helping detect potential bottlenecks and design measures to mitigate them.

2 Conceptual Foundations

This paper investigates a method for describing the input-output balance using a system of ordinary differential equations. W. Leontief first proposed this method in 1952 [7]. The model parameters are: B – the matrix of capital coefficients, A – the matrix of direct input coefficients, $X(t)$ – the vector function of gross output, and VC – the vector of non-productive consumption

$$X(t) = B \cdot \frac{dX(t)}{dt} + A \cdot X(t) + VC(t),$$

or if we use the matrix of non-productive consumption coefficients C, then $VC(t) = C \cdot X(t)$, and the equation takes the form

$$X(t) = B \cdot \frac{dX(t)}{dt} + A \cdot X(t) + C \cdot X(t), \quad X(0) = X_0. \tag{1}$$

Let's consider transforming the system of differential equations into a canonical form, considering that matrices A, B and C are calculated using the formulas:

$$A = MIC \cdot \operatorname{diag}\left(\frac{1}{X_0}\right),$$

$$B = \operatorname{diag}\left(\frac{IV}{X_0}\right) \cdot TIC \cdot \operatorname{diag}(Y) \cdot \operatorname{diag}\left(\frac{\Delta t}{\Delta X}\right),$$

$$C = \left(\left(E - \operatorname{diag}\left(\frac{IV}{X_0}\right)\right) \cdot TIC \cdot \operatorname{diag}(Y) - MIC\right) \cdot \operatorname{diag}\left(\frac{1}{X_0}\right),$$

where IV is the investment vector, MIC is the matrix of intermediate consumption, Y is the vector of final demand, $TIC = (E - A)^{-1}$ is the matrix of total input coefficients (E being the identity matrix of the corresponding order), $\Delta X = X_{end} - X_0$, where X_{end} is the target vector of gross outputs after a time period Δt, and X_0 is the initial (current) vector of gross outputs.

Important note: Throughout this work, the notation $\operatorname{diag}\left(\frac{\alpha}{\beta}\right) = \begin{pmatrix} \frac{\alpha_1}{\beta_1} & 0 & \cdots & 0 \\ 0 & \frac{\alpha_2}{\beta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\alpha_n}{\beta_n} \end{pmatrix}$ $\operatorname{diag}\left(\frac{c}{\beta}\right) = \begin{pmatrix} \frac{c}{\beta_1} & 0 & \cdots & 0 \\ 0 & \frac{c}{\beta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{c}{\beta_n} \end{pmatrix}$ for vectors $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ and a constant c.

Expression (1) can be reduced to the Cauchy normal form:

$$\frac{dX(t)}{dt} = GX(t), \quad G = B^{-1}(FFC - A - C), \quad X(0) = X_0, \tag{2}$$

where $FFC = TIC \cdot \operatorname{diag}(Y) \cdot \operatorname{diag}\left(\frac{1}{X_0}\right)$ is the matrix of financial flow coefficients.

The inverse matrix B^{-1} of the singular matrix B is replaced by its pseudoinverse, computed via Singular Value Decomposition (SVD) [13].

Important note: For the correct inversion of matrix B via SVD, it is necessary to permute the rows (and columns) of the original matrices MIC, IV, Y, ΔX and X_0 such that the zero rows (columns) of MIC and IV are positioned at the bottom (right); the rows and columns of the MIC matrix must be permuted accordingly.

The solution to the Cauchy problem (2) is given by the formula

$$X(t) = e^{Gt} \cdot X_0, \tag{3}$$

where e^{Gt} is the matrix exponential.

The series define the matrix exponential

$$e^G = E + G + \frac{G^2}{2!} + \cdots = \sum_{k=0}^{\infty} \frac{G^k}{k!},$$

and this series converges absolutely.

Accordingly, the representation of the matrix exponential e^{Gt} can be derived from the expression

$$e^{Gt} = E + Gt + \frac{G^2}{2!}t^2 + \cdots = \sum_{k=0}^{\infty} \frac{G^k}{k!}t^k. \tag{4}$$

The pseudocode for obtaining the analytical solution (3) is presented in Algorithm 1.

Algorithm 1. Obtaining the analytical solution to problem (2).
Input: n, Δt, X_0, ΔX, Y, MIC, IV
Output: $e^{G\Delta t}$, $X(\Delta t)$
1: $X_0 = swap(X_0)$
2: $\Delta X = swap(\Delta X)$
3: $Y = swap(Y)$
4: $MIC = swap(MIC)$
5: $IV = swap(IV)$
6: $A = MIC \cdot \ diag\ (1/X_0)$
7: $TIC = (E - A)^{-1}$
8: $B = \ diag\ (IV/X_0) \cdot TIC \cdot \ diag\ (Y) \cdot \ diag\ (\Delta t/\Delta X)$
9: $C = ((E - \ diag\ (IV/X_0)) \cdot TIC \cdot \ diag\ (Y) - MIC) \cdot \ diag\ (1/X_0)$
10: $FFC = TIC \cdot \ diag\ (Y) \cdot \ diag\ (1/X_0)$
11: $B^{-1} = SVD(B)$
12: $G = B^{-1}(FFC - A - C)$
13: $X(t) = e^{Gt} \cdot X_0$
14: **return** $e^{G\Delta t}$, $X(\Delta t)$

The integral curves $X(t)$ of the solution (3) to the Cauchy problem (2) will contain output estimation errors. The sources of error are the absence of forecast deflator values by economic activity type and the pseudoinversion of the inertia matrix. As the forecast horizon increases, the absolute and relative magnitudes of these errors will rise.

We will estimate the forecasting errors of the computed gross output indicators using specific data.

3 Solution of the System of Differential Equations for the Dynamic Input-Output Balance

We demonstrate the features of applying model (1) for economic activity planning and evaluate its accuracy using the example of open data from the symmetric input-output tables of the Russian Federation for 2011 and 2016. The choice of this specific time interval is justified by the fact that, for a long time after 2016, the official input-output tables of the Russian Federation were not published in open access.

The problem can be interpreted as follows: given the initial input-output table data for 2011 and the target (desired) gross output for 2016, it is required to obtain a solution to the system of differential Eqs. (1) that most accurately forecasts the gross production by industry for 2016. Then, guided by the obtained solution, it is possible to regulate inter-industry production and resource consumption to achieve the planned targets.

In solving the formulated problem, particular attention will be paid to the influence of statistical data errors on the accuracy of the solution to differential Eq. (1).

We will compute matrices A, B, and C using the formulas provided in Subsection 1.2 for the problem stated above:

$$A = MIC_{2011} \cdot \mathrm{diag}\left(\frac{1}{X_{2011}}\right),$$

$$B = \operatorname{diag}\left(\frac{IV_{2011}}{X_{2011}}\right) \cdot TIC_{2011} \cdot \operatorname{diag}(Y_{2011}) \cdot \operatorname{diag}\left(\frac{5}{\Delta X}\right),$$

$$C = \left(\left(E - \operatorname{diag}\left(\frac{IV_{2011}}{X_{2011}}\right)\right) \cdot TIC_{2011} \cdot \operatorname{diag}(Y_{2011}) - MIC_{2011}\right) \cdot \operatorname{diag}\left(\frac{1}{X_{2011}}\right),$$

where IV_{2011} is the investment vector (column 108 of the corresponding 2011 symmetric input-output table), MIC_{2011} is the matrix of intermediate consumption (columns 1–98 of the corresponding 2011 symmetric input-output table), Y_{2011} is the vector of final demand (column 110 of the corresponding 2011 symmetric input-output table), $TIC_{2011} = (E - A)^{-1}$ is the matrix of total input coefficients, $\Delta X = X_{2016} - X_{2011}$, and X_{2016} and X_{2011} are the vectors of gross output (column 111 of the symmetric input-output table for 2016 and 2011, respectively).

The Cauchy problem (2) is then formulated as follows:

$$\frac{dX(t)}{dt} = GX(t), \quad G = B^{-1}(FFC_{2011} - A - C), \quad X(0) = X_{2011},$$

where $FFC_{2011} = TIC_{2011} \cdot \operatorname{diag}(Y_{2011}) \cdot \operatorname{diag}\left(\frac{1}{X_{2011}}\right)$ is the matrix of financial flow coefficients.

Proceeding to direct computations, it is first necessary to replace the matrix B^{-1} with the pseudoinverse of B via SVD. Calculating the pseudoinverse matrix does not present significant difficulties, as SVD is implemented in most well-known mathematical software packages and many high-level programming languages.

It is more important to note that, in accordance with the note, the rows (and columns) of the matrices MIC_{2011}, IV_{2011}, Y_{2011}, ΔX, X_{2011} must be permuted such that the zero rows (and columns) of MIC_{2011} and IV_{2011} are positioned at the bottom (right). This transformation is not unique, as there are numerous ways to permute the rows (and columns) to satisfy the stated condition. To ensure the reproducibility of the results obtained in this work, we specify the new order of rows (and columns) for the original matrices, as determined by the author: 70, 52, 60, 89, 61, 72, 90, 9, 47, 86, 57, 56, 59, 55, 1, 46, 32, 71, 58, 7, 96, 76, 49, 63, 62, 14, 73, 21, 45, 88, 4, 41, 8, 53, 33, 6, 36, 24, 25, 40, 13, 22, 51, 44, 18, 64, 37, 42, 27, 38, 26, 54, 29, 65, 43, 77, 15, 28, 23, 35, 5, 20, 31, 16, 12, 79, 30, 19, 17, 39, 34, 80, 94, 50, 10, 78, 11, 48, 2, 66, 67, 68, 69, 74, 75, 81, 82, 83, 84, 85, 87, 91, 92, 93, 97, 3, 95, 98.

The presence of 20 zeros in the IV vector of the symmetric input-output table, resulting from differences in the nomenclature of the 2011 and 2016 tables or the absence of data, leads to the appearance of zero rows in the matrix B. Consequently, model (1) will comprise 78 differential and 20 algebraic equations. Thus, it is impossible to forecast the gross output for these 20 industries after the time interval Δt using model (1) alone (for the simple reason that the components of the solution corresponding to these industries will not depend on time t). Therefore, the forecast quality will subsequently be evaluated only for the 78 industries for which such a forecast is possible.

The solution (3) to the formulated Cauchy problem is given by the formula

$$X(t) = e^{Gt} \cdot X_{2011}.$$

Having constructed the matrix exponential for $\Delta t = 2016 - 2011 = 5$ years, we obtain the computed solution $\tilde{X}_{2016}$ for the gross output indicators by industry for the year 2016.

$$\tilde{X}_{2016} = X(\Delta t = 5) = e^{5G} \cdot X_{2011} \tag{5}$$

The computation of the matrix exponential e^{5G}, like SVD, is implemented in most well-known mathematical software packages and many high-level programming languages.

We will estimate the Mean Relative Percentage Error (MRPE) of the computed gross output indicators for 2016. When interpreting the results and evaluating the MRPE, it is important to remember that the solution $\tilde{X}_{2016}$ was obtained using the input-output table data whose rows (and columns) were permuted. Therefore, before comparing it with the exact solution X_{2016}, it is necessary to either permute the rows of the particular solution $\tilde{X}_{2016}$ back to their original positions or permute the rows of the actual solution X_{2016} according to the permutation applied to the original 2011 input-output table data.

The MRPE for 2016 across the 78 industries (corresponding to the number of non-zero elements in the *IV* vector) is calculated in the standard manner:

$$MRPE = \frac{1}{78} \sum_{i=1}^{78} \frac{\left|\tilde{X}_{2016_i} - X_{2016_i}\right|}{X_{2016_i}} \cdot 100\%. \tag{6}$$

The calculation of the gross output for 2016 across the 78 categories of the input-output table according to (5) yields an MRPE of 16.72%. Given the chosen medium-term time interval of 5 years, the achieved accuracy can already be considered acceptable; however, this result can be improved by refining the values of the matrix exponential.

4 Neural Network Method for Refining the Solution of the Dynamic Input-Output Balance System of Differential Equations

We formulate an optimization problem to refine the solution (5) obtained from the dynamic input-output balance system of differential equations and solve it using an Artificial Neural Network (ANN) framework.

In the case of zero forecasting error for 2016, the following equality holds:

$$\Delta X = \left(e^{G\Delta t} - E\right) \cdot X_0, \tag{7}$$

however, since an exact solution is unattainable, we formulate the optimization problem as follows:

$$L(W) = \sum_{i=1}^{n} \left(W_i \cdot X_0 - \Delta X_i\right)\Big)^2 \to, \tag{8}$$

where W is the matrix of tunable (trainable) parameters, $W^{(0)} = e^{G\Delta t} - E$ is the initial approximation for the trainable parameters, and $L(W)$, in the terminology of Artificial

Neural Networks (ANN) and Machine Learning (ML), is the expression for estimating the empirical risk or the loss function that must be minimized.

It is proposed to solve the optimization problem (8) using the ANN framework. The formulation of problem (8) is analogous to a regression restoration task. From an architectural standpoint for the practical problem at hand, a fully connected single-layer ANN is initially proposed, with 98 inputs (to which the components of the vector X_{2011} are fed) and, correspondingly, 98 outputs (at which the values of the vector ΔX must be approximated). If the forecast quality is unsatisfactory, the number of hidden layers in the network will be gradually increased.

Since the ANN is fully connected, the initial number of adjustable parameters (network weights) is $98 \times 98 = 9604$. Bias weights are not used because the system of differential equations is already in canonical form (2) (Fig. 1).

The sample size of 98 instances (i.e., the number of components in the vector), each characterized by 98 parameters, allows the use of full gradient descent to adjust the ANN weights (i.e., train the ANN).

As the initial approximation $W^{(0)}$, we use the matrix exponential obtained for the solution $\tilde{X}_{2016}$, minus the identity matrix, i.e., $e^{5G} - E$.

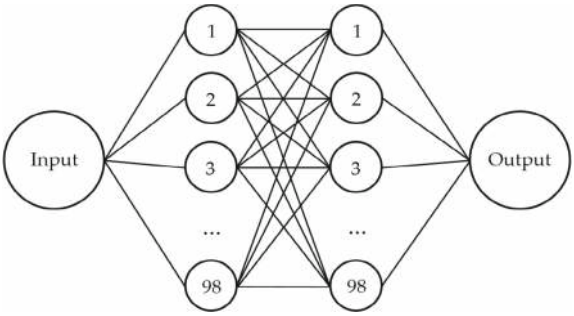


Fig. 1. The initial ANN architecture for refining the analytical solution (5) of the dynamic input-output balance system of differential equations.

We will compute the matrix $dL(W)$ of the following form to implement gradient descent:

$$dL(W) = \left(\frac{\partial L(W)}{\partial W_{11}} \; \frac{\partial L(W)}{\partial W_{12}} \cdots \frac{\partial L(W)}{\partial W_{1n}} \; \frac{\partial L(W)}{\partial W_{21}} \; \frac{\partial L(W)}{\partial W_{22}} \cdots \frac{\partial L(W)}{\partial W_{2n}} \; \vdots \; \ddots \; \vdots \; \frac{\partial L(W)}{\partial W_{n1}} \; \frac{\partial L(W)}{\partial W_{n2}} \cdots \frac{\partial L(W)}{\partial W_{nn}} \right),$$

where $n = 98$.

The iterative process for refining the values of the parameter matrix (i.e., training the ANN) is carried out according to the rule:

$$W_{ij}^{(k+1)} = W_{ij}^{(k)} - h \cdot dL(W)_{ij}\Big|_{W_{i1}=W_{i1}^{(k)}, W_{i2}=W_{i2}^{(k)}, \ldots, W_{in}=W_{in}^{(k)}}, \tag{9}$$

where $i = 1..n, j = 1..n, n = 98$, and h is the gradient step (a value of $h = 10^{-16}$ was used during training).

It should be noted that, for the problem at hand, monitoring changes in forecast quality via the empirical risk function is not particularly convenient (although this metric will, of course, be used during training). It is more illustrative to track changes in forecast quality by monitoring the MRPE averaged across the 78 industries for 2016. This will require the construction of a refined solution $\tilde{X}_{2016}^{(k)}$: at each k-th iteration:

$$\tilde{X}_{2016}^{(k)} = \left(W^{(k)} + E\right) \cdot X_{2011} \tag{10}$$

The formula for calculating the MRPE for 2016 across the 78 industries (denoted as n_0) at the k-th iteration is as follows:

$$MRPE^{(k)} = \frac{1}{78} \sum_{i=1}^{78} \frac{\left|\tilde{X}_{2016_i}^{(k)} - X_{2016_i}\right|}{X_{2016_i}} \cdot 100\%.$$

The training process terminates upon reaching the required $MPRE_{threshold}$ value or after a specified number of full gradient descent iterations, $epoch_{max}$, set by the researcher. In the numerical example presented in this work, the values $MPRE_{threshold} = 0.001\%$ and $epoch_{max} = 500$ were used.

The pseudocode for the proposed neural network method for refining the analytical solution (3) is presented in Algorithm 2.

Algorithm 2. Neural network method for refining the analytical solution to problem (2).

Input: n, n_0, $e^{G\Delta t}$, X_0, X_{end}, ΔX, h, $MPRE_{threshold}$, $epoch_{max}$
Output: W

1: $L(W) = \sum_{i=1}^{n} \left(W_i \cdot X_0 - \Delta X_i\right)\Big)^2$
2: $dL(W) = \{\partial L(W)/\partial W_{ij}\}$, $i, j = 1..n$
3: $k = 0$
4: $W^{(k)} = e^{G\Delta t} - E$
5: $\tilde{X}{}_{end}^{(k)} = \left(W^{(k)} + E\right) \cdot X_0$
6: $MRPE^{(k)} = \sum_{i=1}^{n_0} \left|\tilde{X}{}_{end_i}^{(k)} - X_{end_i}\right| / X_{end_i} / n_0 \cdot 100$
7: **while** $MRPE^{(k)} > MPRE_{threshold}$ **and** $k < epoch_{max}$ **do**
8: $W_{ij}^{(k+1)} = W_{ij}^{(k)} - h \cdot dL(W)_{ij}\Big|_{W_{i1}=W_{i1}^{(k)}, W_{i2}=W_{i2}^{(k)}, \ldots, W_{in}=W_{in}^{(k)}}$, $i, j = 1..n$
9: $k = k + 1$
10: $\tilde{X}{}_{end}^{(k)} = \left(W^{(k)} + E\right) \cdot X_0$
11: $MRPE^{(k)} = \sum_{i=1}^{n_0} \left|\tilde{X}{}_{end_i}^{(k)} - X_{end_i}\right| / X_{end_i} / n_0 \cdot 100$
12: **return** $W^{(k)} + E, \tilde{X}{}_{end}^{(k)}$

Algorithm 2, similarly to Algorithm 1, returns a refined value of the matrix exponential and its corresponding approximate solution at $t = \Delta t$.

5 Results

The dependence of the MRPE for 2016 on the number of full gradient-descent iterations, is shown in Fig. 2.

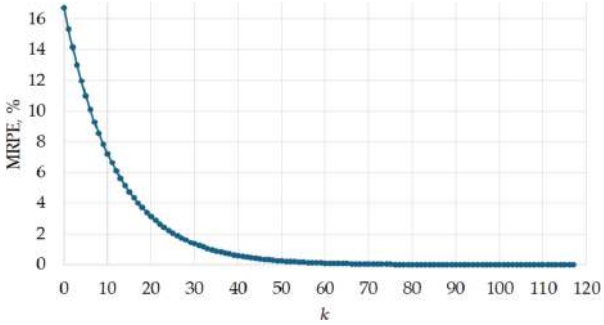


Fig. 2. Dependence of the 2016 MRPE on the number of full gradient descent iterations.

Refining the particular solution $\tilde{X}_{2016}$ of the Cauchy problem in the normal form (2) – obtained analytically and representing the forecast of gross output for 78 industries in 2016 – using ML and ANN methods, reduced the MRPE from 16.72% to 0.001% over 117 iterations of full gradient descent (or 117 training epochs in ANN terminology) without increasing the architecture of the fully connected single-layer ANN (Fig. 1). The refinement process can be continued. The solution can be further improved, given sufficient computational resources and a correct choice of the gradient step *h*.

6 Discussion

Why is there no overfitting control? Typically, ML and ANN imply splitting the entire sample of objects into training and test sets. This is done to control overfitting, i.e., to avoid overfitting to the training data. Translating this circumstance to the applied problem being solved, it can be noted that the optimization problem was formulated precisely in such a way as to most closely approximate the forecast of gross output by economic sector, rather than to construct a solution with generalizing capability.

Of course, it will not be possible to infinitely refine the analytical solution by engaging ever-greater computational resources and expanding the ANN architecture. Why? Substantial absolute weights characterize the symptomatology of overfitting. Interpreting the obtained solution, the weights indicate which indicators in the original input-output table should be increased and which should be reduced to achieve the planned gross output after a selected time interval. Accordingly, large absolute weights indicate significant changes in production and resource consumption across specific sectors. This is, firstly, not always possible for objective reasons (a considerable restructuring of an industry, in terms of production/consumption of resources, requires a lot of time) and, secondly, can lead to an undesirable inter-industry imbalance (when the most efficient sector in terms of production/consumption ratio is maximally loaded, while other sectors

are idle). Therefore, weight control (and, consequently, overfitting control) should be carried out by the researcher based on the chosen economic strategy, developed economic planning, and considering the flexibility of the country's/region's economy (what volume of changes specific industries are prepared for and how quickly they can be achieved), rather than by the standard approach with a test set.

Note that the proposed neural network method refines forecasts across all industries, not only those for which an analytical solution is available, thereby overcoming one of the main drawbacks of the dynamic input-output balance model (1).

It is critical to note and explain that achieving such accuracy without extensive computational resources in an acceptable time was enabled primarily by the initial analytical approximation. The formulated problem (8) is characterized by many tunable parameters (9604) and a small sample size (98 instances), making it multi-extremal. In this situation, the importance of the initial approximation becomes critical. A poorly chosen initial approximation can, at best, significantly slow down the convergence of the gradient descent to a minimum; and at worst (which is also the most probable), lead to convergence not to the global minimum, but only to a local one. The probability of such a scenario increases substantially when solving a multi-extremal problem. Therefore, the high-quality initial approximation, obtained analytically, enabled the achievement of such an accurate result.

7 Conclusion

The input-output balance is a crucial tool for economic analysis and forecasting, enabling the identification of potential issues and the implementation of targeted solutions. Describing the input-output balance through a system of ordinary differential equations represents the most adequate approach for digitizing this model. Its analytical solution yields sufficiently accurate short-and medium-term forecasts; however, the resulting solutions can be further refined.

The corresponding optimization problem formulated in this study, together with the proposed neural network method for its solution, reduced the MRPE from 16.72% to 0.001% for a medium-term 5-year forecast of gross output by economic sector in the Russian Federation for 2016, based on data from the 2011 symmetric input-output tables. It is particularly noteworthy that this result was achieved in just 117 iterations of full gradient descent (or 117 training epochs in ANN terminology) using the simplest architecture of a fully connected single-layer ANN with 98 inputs and 98 outputs. Achieving such precision without extensive computational resources within an acceptable timeframe was enabled mainly by the high-quality initial approximation obtained analytically.

Acknowledgments. The research was carried out at the expense of a grant from the Russian Science Foundation no. 24-28-00056 (https://rscf.ru/project/24-28-00056/).

References

1. Baimenov, A.M.: About some of the factors for public administration effectiveness. Public service. **123**(1), 26–32 (2020) https://doi.org/10.22394/2070-8378-2020-22-1-26-32

2. Bašić, D., Ćurić, P.: Adaptability of securitization model to conditions of volatile financial structure. Economics. **9**(1), 205–219 (2021) https://doi.org/10.2478/eoik-2021-0012
3. Ganic, M.: Are determinants of international financial integration in the European transition countries different from post-transition countries? SSRN Electron. J., 3725386 (2020) https://doi.org/10.2139/ssrn.3725386
4. Grigoryev, L., Zharonkina, D.: China's capital formation in the volatile time. BRICS. J. Econ. **4**(2), 265–283 (2023) https://doi.org/10.3897/brics-econ.4.e105980
5. Gusev, V.B.: Binary model of restructuring the technology complex of economy. In: Proceedings of the 2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), pp. 13–17. Lipetsk, Russian Federation (2022) https://doi.org/10.1109/SUMMA57301.2022.9973918
6. Nemchinov, V.S.: Social Value and Planned Price; Science: Moscow, USSR, 1970. 507
7. Toroptsev, E.L.; Kandokhova, M.M.; Gudieva, N.G.: Differential equations of input–output balance: digitization and the Prospect of application. Econ. Contemp. Russ. 2024, **1**, 7–22. https://doi.org/10.33293/1609-1442-2024-1(104)-7–22
8. Ganic, M.: Financial Globalization in the Emerging Balkans Exploring Financial Trends on the eve of EU Membership, p. 151. Palgrave Macmillan, Cham, Switzerland (2020) https://doi.org/10.1007/978-3-030-65009-4
9. Ristic, K.: Meta analyses and controversy about Financialization, growth and stability: a look from post-transition countries of last decade. Economics. **8**(1), 41–49 (2020) https://doi.org/10.2478/eoik-2020-0005
10. Rusek, A.: Financial integration and the new EU member countries: challenges and dilemmas. Prague Economic Papers 2005, 14(1), 17–32. https://doi.org/10.18267/j.pep.250
11. Voslensky, M.: Nomenklatura: the Soviet Ruling Class, p. 455. New York City, United States, Doubleday (1984)
12. Yucis, A.E. : Problems of the strategy of sustainable development of manufacturing enterprises. Scientific Research and Development. Economics 2021, **9**(3), 33–36. https://doi.org/10.12737/2587-9111-2021-9-3-33-36
13. Zhang, Z.: The Singular Value Decomposition, Applications and beyond. (2015). https://doi.org/10.48550/arXiv.1510.08532

Feature Space Optimization and Ensemble Learning Models for Robust Network Intrusion Detection

Mohammad Hashir[1], Mohammad Zayeem[1(✉)], Maria Lapina[2], and Mohammad Sajid[1]

[1] Aligarh Muslim University, Aligarh 202002, India
zayeemazmi@gmail.com
[2] North Caucasus Federal University, 355017 Stavropol, Russia

Abstract. The increasing challenge of network traffic and cyberattacks renders the use of traditional intrusion detection systems (IDS) ineffective on high-dimensional and imbalanced data. To address these weaknesses, this paper proposes a hybrid IDS model that integrates metaheuristic-based feature selection models with an ensemble learning models. Five optimization algorithms, namely Particle Swarm Optimization (PSO), Sine Cosine Algorithm (SCA), Fungal Growth Optimization (FGO), Cuckoo Catfish Optimizer (CCO), and Fossa Optimization Algorithm (FOA), are combined with XGBoost, LightGBM, CatBoost, Soft Voting, and Stacking ensembles. The top two optimizers are selected by accuracy, their features are fused, and a stacked ensemble is trained on the fused set. The experiments conducted on the UNSW-NB15, NSL-KDD, and KDDCup99 datasets indicate that metaheuristic optimization is a suitable approach for optimizing detection accuracy with reduced feature dimensionality. Overall, the proposed hybrid intrusion detection framework demonstrates consistent and reliable performance across multiple datasets, making it suitable for modern and evolving network environments.

Keywords: Feature selection · Metaheuristic algorithms · Machine learning · UNSWB-N15 · NSL_KDD · KDDCup99

1 Introduction

The Internet of Everything (IoE) represents a fundamental transformation in modern society, driven by the digital transformation era, which enables seamless connectivity among people, processes, data, and things [1, 2]. This has led to a shift in cybersecurity from a technical challenge to a strategic requirement for organizations globally [3]. To uphold the CIA principles, multi-layered defense mechanisms, such as firewalls, antivirus software, and Intrusion Detection Systems (IDS), are widely implemented [4]. Among them, the Network Intrusion Detection System (NIDS) is essential due to its real-time network monitoring capabilities, ability to identify malicious behavior [5, 6]. However, recent incidents, such as the Colonial Pipeline ransomware attack [7], have brought implicit trust models in security architecture to the forefront [8]. In conventional NIDS, both signature-based and anomaly-based detection methods are used.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 182–195, 2027.
https://doi.org/10.1007/978-3-032-27834-0_18

Signature-based methods are effective at deterring known threats; however, they are less effective at detecting zero-day threats and new malware variants, resulting in rapid obsolescence [9]. Anomaly-based systems attempt to model the normal behavior of a network but frequently exhibit large false-positive rates and fail to accurately capture complex and dynamic traffic patterns, especially in large networks [9, 10]. These flaws have contributed to the increased adoption of machine learning (ML) and artificial intelligence (AI) in NIDS [11]. Recent research has proven the effectiveness of AI-based optimization and reinforcement learning in solving complex, high-dimensional problems, including the Vehicle Routing Problem [12]. ML-based NIDS utilize the data from considerable network traffic to identify previously unknown attacks and are growing increasingly important in protecting against Advanced Persistent Threats (APTs). The biggest shortcoming is the imbalance of the data because the proportion of malicious traffic in the entire network activity is only a small percentage [13]. This skew makes the learning models biased in favor of normal traffic, and thus reduces their capacity to capture infrequent yet vital cases of attacks, potentially causing security blind spots [14]. Moreover, the scale and the data heterogeneity in distributed environments grow exponentially, making calculations more complex [15, 16]. The use of AI-based optimization to solve the problem of scalability and dynamic resource management in cloud systems has been demonstrated to be effective [17], and adapted metaheuristic methods, including the Fossa Optimization Algorithm, have been proven to help deal with the issues of heterogeneity and large-scale optimization in the conditions of IoT and cloud systems [18]. Even though ML-based NIDS represent a significant breakthrough in cybersecurity, their efficiency is still limited by issues of class imbalance, non-stationarity, and high-dimensional learning conditions. Current solutions are not robust and flexible enough to effectively respond to existing and emerging threats regularly. To eliminate these gaps, this paper proposes a hybrid IDS model that integrates metaheuristic-based feature selection models with an ensemble learning models that aims to overcome the problem of class imbalance by utilizing a more effective data preprocessing mechanism and hybrid algorithmic approaches, thereby more effectively identifying minority attacks without compromising overall system performance. Five optimization algorithms, namely Particle Swarm Optimization (PSO), Sine Cosine Algorithm (SCA), Fungal Growth Optimization (FGO), Cuckoo Catfish Optimizer (CCO), and Fossa Optimization Algorithm (FOA), are combined with XGBoost, LightGBM, CatBoost, Soft Voting, and Stacking ensembles.

2 Related Research

This section provides a review of recent studies, with an emphasis on their methodologies, contributions, and limitations. Lorenzo Diana et al. [19] presented an extensive review of IDS, including its architectures, detection methods, and applications in cloud computing, IoT, and industrial control systems. Saima Farhan et al. [20] suggested ResNet-CNN to address the limitations of traditional IDS, such as manual feature extraction and less generalization. Their model has achieved detection rates of over 98.9% when classifying binary and multiclass data on the NSL-KDD dataset, with detection rates exceeding 99% for individual classes of attacks. Rahma Jablaoui et al. [21] proposed a deep learning

IDS that integrates CNNs to extract spatial features with Recurrent Neural Networks (RNNs), i.e., LSTM, BiLSTM, GRU, and BiGRU, to learn the temporal relationship. Muhammad Farhan et al. [22] suggested a Sequential Deep Neural Network (DNN) using ReLU activation and an Extra Tree Classifier as a feature selection model. Their model reduced the number of features to eight main attributes and maintained a detection rate of 97.93 on the UNSW-NB15 dataset, albeit at a higher computational cost. To overcome the limitations of the IoT environment, Abdullah Mujawib Alashjaee et al. [23] designed a hybrid IDS based on Principal Component Analysis (PCA) to reduce the number of features and a mixed Feed Forward Neural Network (FFNN)-XGBoost classifier. In addition to detection performance, Fatemeh Ebrahimi et al. [24] indicated the absence of interpretability in ML/DL-based IDS. They suggested a one-dimensional, lightweight CNN fused with SHAP and LIME explanation methods, achieving 99.5% accuracy on the TON-IoT dataset and enhancing model transparency as well. The feature selection, which was optimized using the FOA, was demonstrated to increase the detection rate and decrease the dimensionality of features [25]. On the same note, R. Kanimozhi et al. [26] proposed a hybrid Pelican Optimization Algorithm-Stacked Sparse Autoencoder (POA-SSAE) model, which achieved high accuracies on the CICIDS2018 and KDDCup99 datasets. Supongmen Walling et al. [27] provided an extensive overview of IoT-based IDS, including existing methods, datasets, and unresolved problems.

The fact is, despite the impressive progress achieved so far in the sphere of IDS due to AI and ML procedures, several crucial gaps remain in the study. Firstly, it lacks a systematic comparison of metaheuristic optimization algorithms of feature selection in IDS [24, 25]. The relative efficiency of algorithms, such as the PSO, SCA, FGO, CCO, and FOA has not been researched in the diversity of attack scenarios. Second, the available methods tend to rely primarily on feature subsets selected by a single optimizer and do not leverage the synergies of alternative optimization methods. The limitation of the absence of feature fusion mechanisms restricts the construction of more discriminative and robust feature spaces. Third, little is investigated on the combination of multi-level or hierarchical ensemble classification systems, which operate in an optimized feature space. Most works priorities the accuracy of detection over computational efficiency and real-time deployment factors, which are crucial in the practical implementation of NIDS. Finally, the convergence behavior, stability, and dynamics of the optimization of the metaheuristic algorithm for IDS feature selection are poorly studied.

This paper aims to fill the above gaps by proposing a systematic framework that involves the analysis of comparative optimization, feature fusion, ensemble classification, and convergence assessment on recent intrusion detection datasets.

3 Proposed Methodology

This paper proposes a hybrid intrusion detection system, which combines metaheuristic-based feature selection with ensemble learning models to maximize performance in detection and computational efficiency. The methodology comprises five major steps: preprocessing of data, categorical encoding and normalization, feature optimization by means of metaheuristics, ensemble classification, and refining the fusion of features. Figure 1 illustrates the overall workflow of the proposed IDS framework.

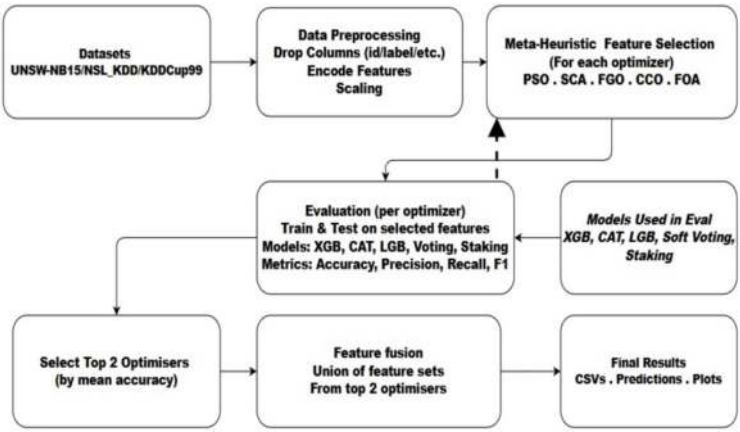


Fig. 1. System workflow for optimizer-guided ensemble intrusion detection.

3.1 Data Preprocessing

Let the dataset be represented as:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \tag{1}$$

where

$$x_i = [x_{i1}, x_{i2}, \ldots, x_{id}] \in \mathbb{R}^d, \quad y_i \in \{0, 1\} \tag{2}$$

Non-informative attributes (id and attack_cat) are removed:

$$X = D \backslash \{\text{id, attack_cat}\} \tag{3}$$

3.1.1 Categorical Encoding

Let

$$C = \{\text{proto, state, service}\} \tag{4}$$

For each $c \in C$, a deterministic mapping is applied:

$$x'_{ic} = f_c(x_{ic}) \tag{5}$$

3.1.2 Feature Normalization

The Z-score standardization is applied:

$$x_{ij}^{norm} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{6}$$

where μ_j and σ_j Denote the mean and standard deviation of the feature. j.

3.2 Metaheuristic-Based Feature Selection

The binary feature selection mask is given as:

$$S = [s_1, s_2, \ldots, s_d], \quad s_j \in \{0, 1\} \tag{7}$$

The reduced feature matrix is:

$$X_S = \left\{ x_{ij} | s_j = 1 \right\} \tag{8}$$

3.2.1 Fitness Function

A multi-objective cost function is used:

$$J(S) = \alpha \cdot E(S) + \beta \cdot \frac{\sum_{j=1}^{d} s_j}{d} \tag{9}$$

The optimal feature set is obtained as:

$$S^* = \arg \min_{S \in \{0,1\}^d} J(S) \tag{10}$$

3.2.2 Metaheuristic Optimization Algorithms

It contains five nature-inspired algorithms, namely, PSO, SCA, FGO, CCO and FOA. Each optimizer iteratively reduces the feature selection vector:

$$S^{(t+1)} = \Phi\left(S^{(t)}, \theta_t\right) \tag{11}$$

where Φ Is the optimiser-specific update function and θ_t Represents dynamic parameters.

3.3 Ensemble-Based Classification

The optimized feature subset X_{S^*} is evaluated using XGBoost, CatBoost, and LightGBM.

3.3.1 Soft Voting Ensemble

$$\hat{y} = \arg\max_k \sum\nolimits_{m=1}^{3} p_m(y = k | X_{S^*}) \tag{12}$$

3.3.2 Stacking Ensemble

$$\hat{y} = f_{\text{meta}}(h_1(X_{S^*}), h_2(X_{S^*}), h_3(X_{S^*})) \tag{13}$$

where f_{meta} is a CatBoost-based meta-learner.

3.4 Feature Fusion Strategy

Let Ω Denote the set of all optimizers and $Acc(\omega)$ The accuracy of the optimizer ω. The top two optimisers are selected:

$$\omega_1, \omega_2 = \arg\max_{\omega \in \Omega} Acc(\omega) \tag{14}$$

The fused feature subset is:

$$S_{\text{fused}} = S_{\omega_1} \cup S_{\omega_2} \tag{15}$$

A final stacked ensemble is trained on $X_{S_{\text{fused}}}$.

4 Results and Discussion

The performance of five metaheuristic feature-selection methods, i.e., PSO, SCA, FGO, CCO, and FOA, is compared and discussed in this section as per their performance on five ensemble learning models, i.e. XGBoost, CatBoost, LightGBM, Soft Voting, and Stacking Ensemble. The performance was evaluated using Accuracy, Precision, Recall and F1-Score [22] and the obtained results are presented in Figs. 2-6.

4.1 Convergence Behavior of Optimization Algorithms

Figure 2 illustrates the convergence of the five metaheuristic algorithms, namely PSO, SCA, FGO, CCO, and FOA, after 50 iterations. PSO was the quickest and most predictable convergence, with the lowest fitness value and early, steady improvement. FOA and CCO came in second and third, respectively, with smooth decreases and minimal fluctuations. FGO exhibited mediocre convergence behavior but needed more iterations to converge, whereas SCA exhibited slow convergence and the most significant final fitness value.

These tendencies indicate that PSO and FOA are more efficient search methods, enabling them to find high-quality subsets of features as quickly as possible. The fact that the convergence curves of PSO and FOA are smoother also indicates that they have a higher capability of exploitation, which is also in line with their high rating in the classification measures that will be presented later.

4.2 Accuracy Analysis

As given in Fig. 3, PSO-XGB provided the most accurate results (~0.992–0.993). FOA-XGB was also another model that performed exceptionally well (~0.992). The ensemble variants (PSO-STACK and FOA-STACK) and stacking were consistently better than the individual models. Such results confirm that PSO and FOA result in an optimal degree of exploration and exploitation, leading to the relevance of pertinent subsets of features.

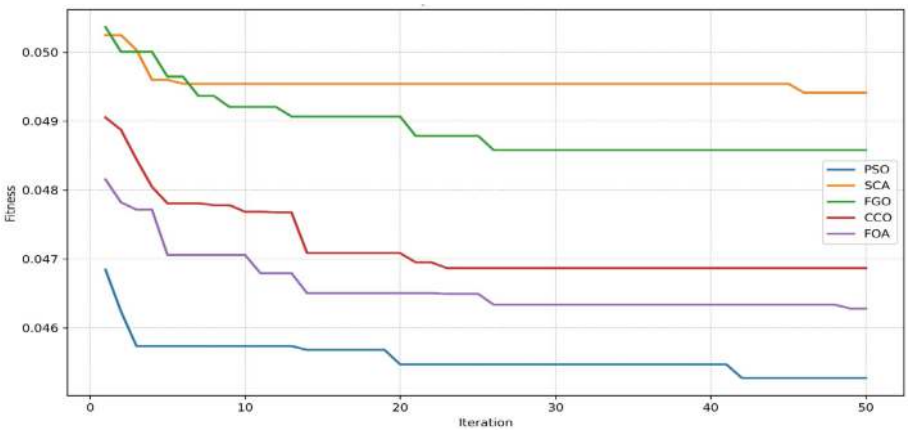


Fig. 2. Convergence curves of metaheuristic algorithms.

4.3 F1-Score Analysis

As shown in Figs. 3 and 4, PSO-XGB and FOA-XGB were once more at the top of the combinations. With the addition of PSO or FOA, F1-scores were slightly higher, indicating a reduced amount of false positives and false negatives. Combination CCOs had the lowest F1-scores (0.982, 0.985). The fact that the F1 values are strong in all the models provides further confirmation of the effectiveness of the selected features in binary intrusion detection.

The highest score in precision was achieved at 0.994–0.996 for the cases of FOA-STACK and PSO-XGB, implying that virtually all the predicted occurrences of attacks were malicious. A combination of SCA and CCO was not that accurate (0.988–0.991). The high levels of precision attest to the fact that the feature subsets effectively reduced false alarms to a minimum, which is essential for cybersecurity systems.

4.4 Precision Analysis & Recall Analysis

As shown in Figs. 5 and 6, PSO-XGB and FOA-XGB were consistently found to provide recall rates of 0.989 or above. CCO-CAT and partial forms of SCA recalled the least (around 0.977 to 0.982), indicating that they missed fewer instances of an attack. However, outcomes are good in the situation of a real-world security dataset.

4.5 Feature Fusion and Final Stacked Model

The top two optimizers in performance on the average performance of all the optimizers were: PSO & FOA. A fused feature subset was created using selected features from PSO and FOA. The final stacked ensemble, which was trained on these combined features, achieved a final Accuracy: ~99%. This confirms that it is possible to enhance robustness by using hybrid feature-selection behaviors through the fusion of two or more optimizers, as these behaviors are complementary.

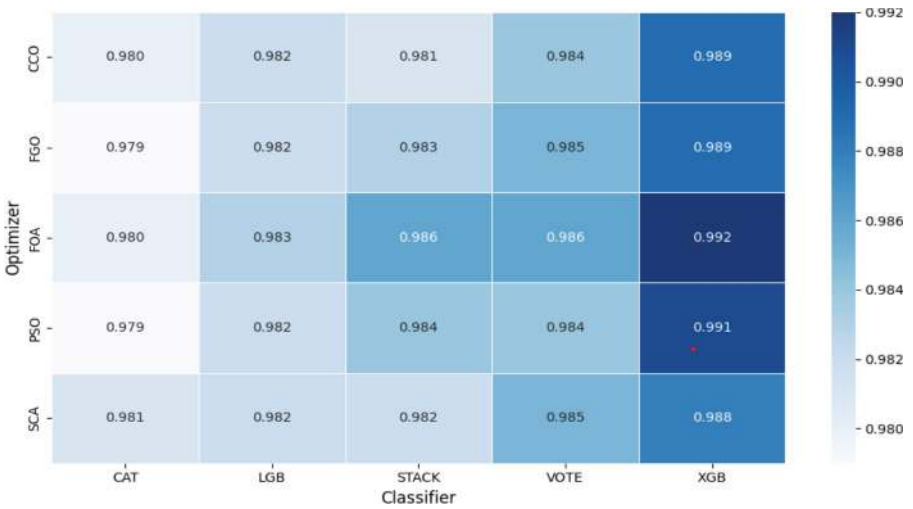


Fig. 3. Accuracy comparison of all optimizer–classifier Pairs.

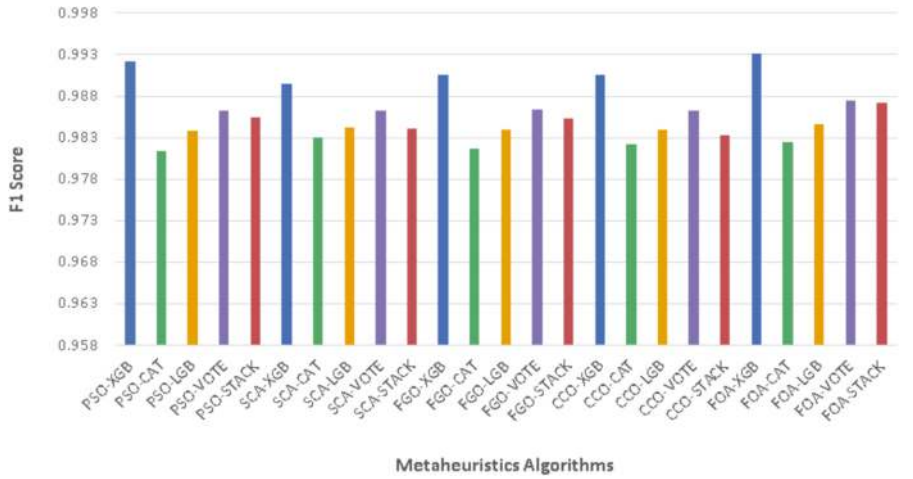


Fig. 4. F1-score comparison of all optimizer–classifier Pairs.

4.6 Observations

Metaheuristics assist in feature selection, yielding a considerable increase in classifier performance. PSO and FOA turned out to be the best optimization solutions of the UNSW-NB15 dataset. The extremely high precision, recall and F1 (0.98) rates indicate high generalization potential and reduced overfitting. There were also PSO/FOA fusion traits that made the models reliable, demonstrating that search strategies can be used to provide more discriminative input subsets. Overall, the paper supports the claim that a hybrid metaheuristic feature selection/ensemble learning configuration is a highly

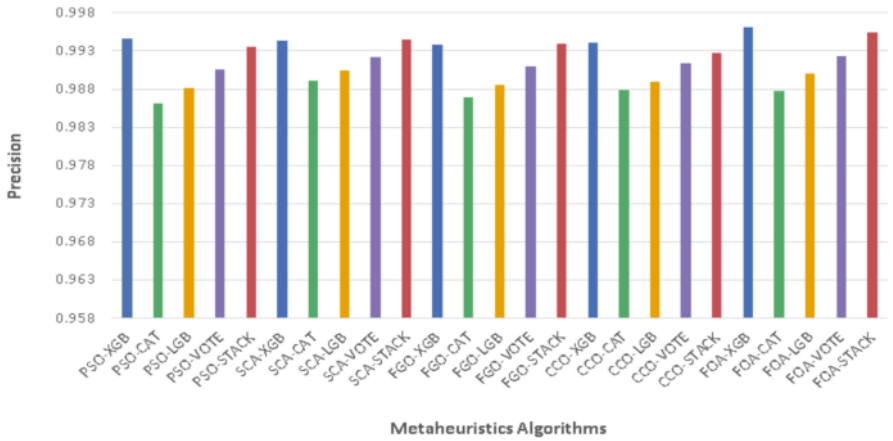


Fig. 5. Precision analysis of all optimizer–classifier Pairs.

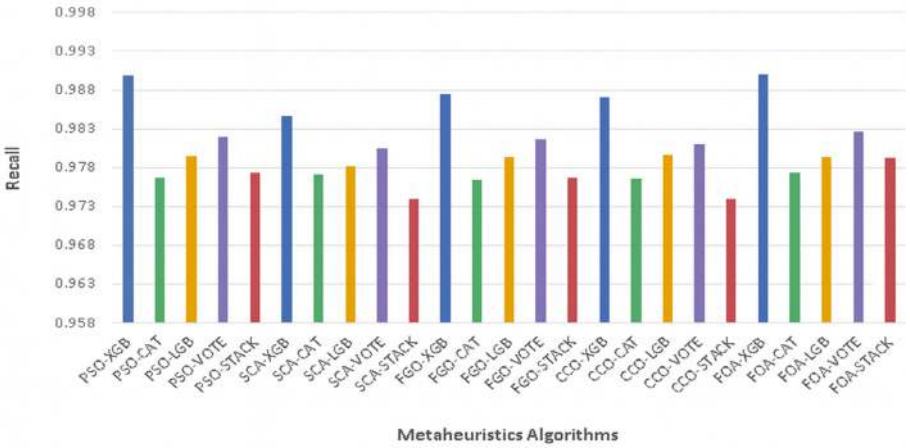


Fig. 6. Recall performance comparison of all optimizer–classifier Pairs.

effective intrusion detection pipeline, with the state-of-the-art performance on UNSW-NB15.

The highest outputs were achieved with the use of PSO-XGB, PSO-STACK, FOA-XGB and FOA-STACK, and the Accuracy and F1-score were more than 0.99 in some instances. It means that the selection of PSO and FOA on UNSW-NB15 data is high in terms of discrimination. CCO-CAT, CCO-LGB and SCA-VOTE, on the other hand, were less competitive and were relatively lower at approximately (0.985–0.988), and this proves that all the optimizers can extract significant subsets of features, but they vary in terms of efficiency.

5 Comparative Analysis Across Multiple Datasets

To demonstrate the robustness and integrity of the hybrid intrusion detection structure, we also compared our experiments with two other benchmark datasets, NSL-KDD and KDDCup99. The methodology has been chosen in the same way: metaheuristic feature selection (PSO, SCA, FGO, CCO, FOA) and ensemble classifiers (XGB, CAT, LGB, VOTE, STACK). This section entails a comparative performance analysis of the performances across all three datasets: UNSW-NB15 (primary), NSL-KDD, and KDDCup99.

5.1 Convergence Behavior Across Multiple Datasets

Data sets were used to study the convergence curves of metaheuristics, as shown in Figs. 7 and 8. Although the convergence rates of PSO and FOA were always rapid and steady, a dataset characteristic influenced both the convergence rate and the final fitness.

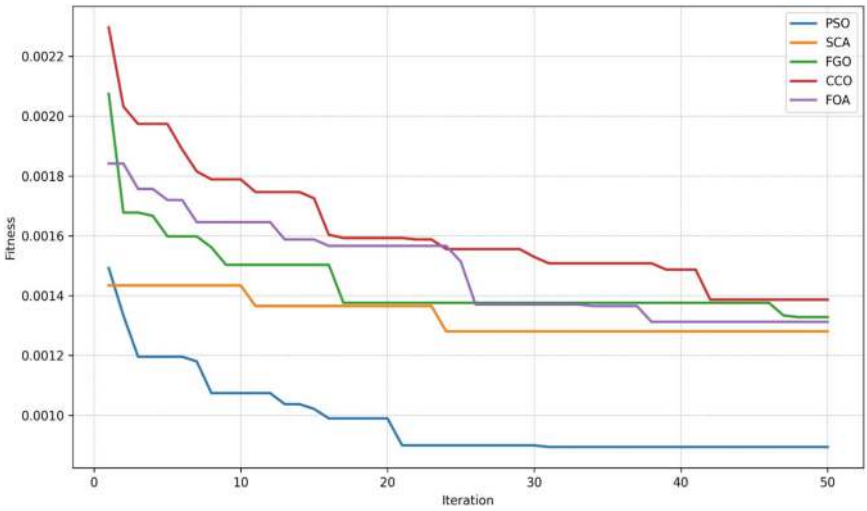


Fig. 7. Convergence curves of the NSL_KDD metaheuristic algorithm.

5.2 Performance of FOA–XGB Across Multiple Datasets

To provide a stable and objective comparison, the outcomes of a hybrid model, Fossa Optimization Algorithm (FOA-XGB), are compared on three benchmark datasets in intrusion detection, including UNSW-NB15, NSL-KDD, and KDDCup99.

According to Table 1, FOA-XGB consistently attained high detection rates across all datasets, regardless of their varying traffic characteristics and attack distributions. The model demonstrated an accuracy of 99.24% and an F1-score of 99.31 on UNSW-NB15, showcasing its strong discriminative abilities in handling complex and contemporary

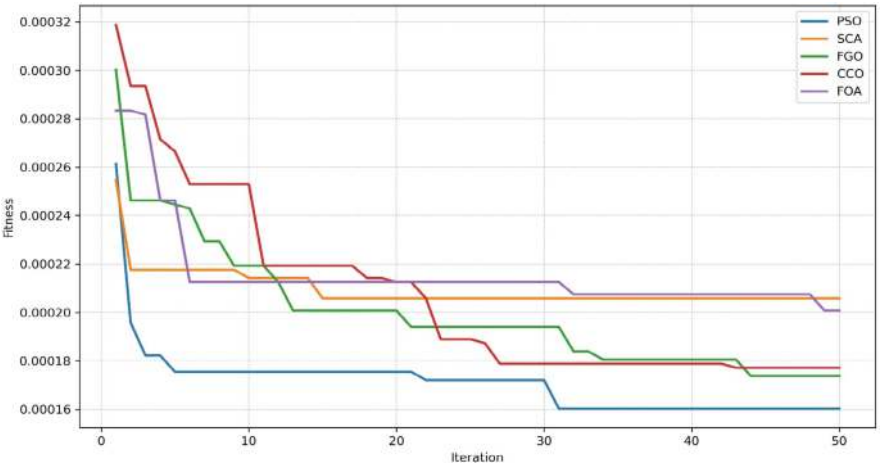


Fig. 8. KDDCup99 convergence curves with metaheuristics.

network traffic. In the NSL_KDD dataset, which is quite sensitive to redundancy and class imbalance, FOA-XGB still performed well, achieving an accuracy of 94.48% and an F1-score of 94.21%.

The most impressive results were obtained on the KDDCup99 dataset using FOA-XGB, which achieved nearly flawless accuracy and F1 scores of 99.96% and 99.97%, respectively. The results indicate a connection between FOA-driven feature selection and XGBoost classification, emphasizing its potential for effective intrusion detection with diverse data, thus underlining its robustness.

Table 1. FOA-XGB model performance on three intrusion detection datasets.

Dataset	Optimiser	Classifier	Accuracy	F1-Score	Selected Features
UNSW-NB15	FOA	XGB	0.9924	0.9931	25
NSL_KDD	FOA	XGB	0.9448	0.9421	23
KDDCup99	FOA	XGB	0.9996	0.9997	21

5.3 Top Performers across Multiple Datasets

The results of the experiment substantiate the fact that the developed framework guarantees a high degree of detection for various types of network intrusion data, as presented in Table 2. The Key observations include:

- PSO and FOA were the best-performing optimisers using all datasets, achieving the best accuracy and F1-scores.
- XGBoost and Voting/Stacking ensembles failed to lose to single classifiers, which demonstrates the effectiveness of ensemble classifiers in intrusion detection.

- This was further improved to Fusion Strategy (they were combining the top-two optimiser feature subsets), which made it more robust, especially in the NSL_KDD and KDDCup99.

Table 2. The best model on three datasets of intrusion detection.

Dataset	Best Optimiser	Best Classifier	Test Accuracy	F1-Score	Selected Features
UNSW-NB15	FOA	XGB	0.9924	0.9931	25
NSL_KDD	CCO	VOTE	0.9461	0.9436	22
KDDCup99	PSO	XGB	0.9998	0.9999	20

5.4 Strength of Feature Selector

The set of features chosen varied across datasets (20–28), and the optimization algorithms attempted to identify the appropriate sets of features, with no overfitting. The size of the feature space was smaller (20 to 22 features) in KDDCup99, allowing for almost perfect performance. The NSL-KDD and UNSW-NB15 datasets required additional features (20–28), which implies the complexity of the network traffic data and the dimensions they currently possess.

5.5 Generalization and Practical Relevance

The viability of the proposed approach is demonstrated by the high scores found in three sets of data. This is particularly crucial in real-world applications, where an intrusion detection system needs to be able to adjust to changing network conditions and assault patterns. When combined with metaheuristic optimisation, the ensemble learning approach appears to be a robust and scalable technique that can ensure accurate detection and reduce false alarms on a wide range of datasets.

6 Conclusion

In this research, an intrusion detection system that combines ensemble learning models with metaheuristic feature selection is proposed and validated. Experiments on three benchmark datasets—NSL-KDD, KDDCup99, and UNSW-NB15 - show that the framework consistently delivers high F1-scores, accuracy, precision, and recall. PSO and FOA provided the most optimistic estimators, while XGBoost and ensemble algorithms demonstrated superiority in classification. The cross-dataset validation confirms the robustness and versatility of the framework. It can be considered to be a feasible framework for integration into an existing network intrusion detection system that requires reliability, scalability, and real-time applicability. Some of the future research directions are as follows-

- The feasible application of the framework in practical network settings to gauge latency and throughput.
- Raw traffic analysis, Deep learning functionality (e.g., CNNs, Transformers).
- Adversarial robustness testing is performed to test resilience to evasion attacks.
- Hybrid optimisation techniques, inlcuding Reinforcmenet Learning should be encouraged for application in IDS.

References

1. Rajadurai, H., Gandhi, U.D.: An empirical model in intrusion detection systems using principal component analysis and deep learning models. Comput. Intell. **37**(3) (2021). https://doi.org/10.1111/coin.12342
2. Evans, D.: The internet of everything: how more relevant and valuable connections will change the world. Cisco IBSG. **2012**(2012), 1–9 (2012)
3. Aroms, E.: NIST special publication 800-94 guide to intrusion detection and prevention systems (IDPS). CreateSpace. (2012)
4. Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., Ahmad, F.: Network intrusion detection system: a systematic study of machine learning and deep learning approaches. Trans. Emerg. Telecommun. Technol. **32**(1) (2021). https://doi.org/10.1002/ett.4150
5. Zhao, X., Fok, K.W., Thing, V.L.L.: Enhancing network intrusion detection performance using generative adversarial networks. Comput. Secur. **145**, 104005 (2024). https://doi.org/10.1016/j.cose.2024.104005
6. Abdulganiyu, O.H., Ait Tchakoucht, T., Saheed, Y.K.: A systematic literature review for network intrusion detection system (IDS). Int. J. Inf. Secur. **22**, 1125–1162 (2023)
7. Beerman, J., Berent, D., Falter, Z., Bhunia, S.: A review of colonial pipeline ransomware attack. In: Proceedings of the 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops, pp. 8–15. IEEE (2023)
8. Ridhawi, I.A., Aloqaily, M.: Zero-trust UAV-enabled and DT-supported 6G networks. In: Proc. IEEE Global Commun. Conf, (GLOBECOM), pp. 6171–6176 (2023)
9. Vitorino, J., Praca, I., Maia, E.: SoK: realistic adversarial attacks and defences for intelligent network intrusion detection. Comput. Secur. **134**, 103433 (2023)
10. Chiew, K.L., Hui, B.: An improved network intrusion detection method based on CNN-LSTM-SA. J. Adv. Res. Appl. Sci. Eng. Technol. **44**(1), 225–238 (2024). https://doi.org/10.37934/araset.44.1.225238
11. Liu, G., Zhang, J.: CNID: research on network intrusion detection based on a convolutional neural network. Discret. Dyn. Nat. Soc. **2020**, 1–11 (2020)
12. Raza, S.M., Sajid, M., Singh, J.: Vehicle routing problem using reinforcement learning: recent advancements. In: Proceedings of the 3rd International Conference on Machine Intelligence and Signal Processing (MISP). NIT Arunachal Pradesh, India (2021)
13. Toldinas, J., Venckauskas, A., Damaevicius, R., Grigaliunas, N.M., Baranauskas, E.: A novel approach for network intrusion detection using multistage deep learning image recognition. Electronics. **10**(15), 1854 (2021)
14. Siddiqi, M.A., Pak, W.: An optimised and hybrid framework for image processing-based network intrusion detection system. Comput. Mater. Contin. **73**(2), 3921–3949 (2022)
15. Chinnasamy, R., Malliga, S., Sengupta, N.: Deep learning-driven intrusion detection systems for smart cities-a systematic study. In: 6th Smart Cities Symposium (SCS 2022), pp. 79–84 (2022)
16. Ashiku, L., Dagli, C.: Network intrusion detection system using deep learning. Procedia Comput. Sci. **185**, 239–247 (2021)

17. Namdev, K., Rajak, R., Sajid, M., Rajak, N.: The role of AI-driven technologies in cloud workflow scheduling: a structured review. J. Ambient. Intell. Humaniz. Comput. (2025). https://doi.org/10.1007/s12652-025-05016-9
18. Qasim, M., Sajid, M.: An efficient IoT task scheduling algorithm in cloud environment using modified firefly algorithm. Int. J. Inf. Technol. **17**(1), 179–188 (2024). https://doi.org/10.1007/S41870-024-017585/FIG.S/5
19. Diana, L., Dini, P., Paolini, D.: Overview on intrusion detection Systems for Computers Networking Security. Computers. **14**(3), 87 (2025)
20. Farhan, S., Mubashir, J., Haq, Y.U., Mahmood, T., Rehman, A.: Enhancing network security: an intrusion detection system using residual network-based convolutional neural network. Clust. Comput. (2025). https://doi.org/10.1007/s10586-025-05156-9
21. Jablaoui, R., Liouane, N.: Network security based combined CNN–RNN models for IoT intrusion detection system. Peer Peer Netw Appl. **18**, 129 (2025) https://link.springer.com/article/10.1007/s12083-025-01944-7
22. Network-based intrusion detection using deep learning technique (2025). https://www.nature.com/articles/s41598-025-08770-0
23. Alashjaee, A.M., Alqahtani, F.: Enhanced intrusion detection system IoT network security model by feed forward neural network and machine learning. Sci. Rep. **15**, 36085 (2025) https://www.nature.com/articles/s41598-025-20047-0
24. Ebrahimi, F., Javidan, R., Akbari, R., Hosseini, Y., et al.: Intrusion detection in the internet of things using convolutional neural networks: an explainable AI approach. Cybersecurity. **8**, 66 (2025). https://doi.org/10.1186/s42400-025-00369-2. https://cybersecurity.springeropen.com/articles/10.1186/s42400-025-00369-2
25. Zayeem, M., Sajid, M.: SFOA-driven defense: leveraging SFOA for optimal feature selection in network intrusion detection systems. In: Proc. Int. Workshop on Advanced (2025)
26. Kanimozhi, R., Neela Madheswari, A.: Novel pelican optimization algorithm (POA) with stacked sparse autoencoder (SSAE) based IDS for network security. Trans. Emerg. Telecommun. Technol. (ETT). (2025). https://doi.org/10.1002/ett.70113. https://onlinelibrary.wiley.com/doi/10.1002/ett.70113
27. Walling, S., Lodh, S.: An extensive review of machine learning and deep learning techniques on network intrusion detection for IoT. Trans. Emerg. Telecommun. Technol. (2025) DOI/publisher page available
28. Sajid, M. et al.: A novel algorithm for capacitated vehicle routing problem for smart cities. Symmetry. **13**(10), 1923 (2021). https://doi.org/10.3390/sym13101923

Comparative Analysis of URL and Hyperlink Features for Machine Learning-Based Phishing Website Detection

Jenefa J[1(✉)], Mary Anita E A[1], Vinodha D[1], and Vladimir Lapin[2]

[1] CHRIST University, Bangalore, India
jenefa.j@christuniversity.in
[2] North Caucasus Federal University, Stavropol, Russia

Abstract. Phishing websites, a type of cyberattack, have become a significant problem in recent years due to the increased access to the internet and computers. In a phishing attack, the attacker tricks people into divulging sensitive information, such as credit card details, personal information, and passwords. The attacker can trick victims, as the website appears similar to a real one and is difficult to differentiate from the real website. There were old methods to detect phishing websites, but they have become ineffective as attackers have evolved, and with that, phishing websites have also evolved. With the advancement in machine learning, a robust system has been developed to detect phishing websites, as it checks for patterns and characteristics that indicate phishing websites. It has proven to be very effective at detecting phishing websites. Most machine learning models utilize URL, content, and network-based features; however, extracting full content-based and network-based features is time-consuming and computationally expensive. In this paper, we compare two feature sets—URL-only and URL + hyperlink features—to assess their performance in phishing detection. The goal is to determine whether adding hyperlink features improves accuracy compared to URL-only features. The dataset was extracted from the Kaggle website, and the models used for this work were SVM, Random Forest, XGBoost, and Stacking Classifier. These models were trained on both feature sets separately, and their performance was evaluated.

Keywords: Phishing attacks · Random forest · Stacking classifier · SVM · XGBoost

1 Introduction

Due to the increased use of the internet, there has been a drastic rise in phishing attacks worldwide. Phishing attacks are critical because they capture sensitive information from users through fake websites that appear genuine, which can be used for illegal activities. According to a recent report from APWG (Anti-Phishing Working Group), there were 1,130,393 phishing attacks in Q2 2025, surpassing the number in Q1 2025 [1]. The numbers are also higher when compared to last year. This indicates an increase in phishing attacks and a corresponding decline in awareness about phishing attacks.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 196–204, 2027.
https://doi.org/10.1007/978-3-032-27834-0_19

Nowadays, it's difficult to identify phishing websites because they look very genuine compared to the original websites. Detecting phishing websites using traditional methods is challenging due to the intricate nature of phishing attacks [2]. The conventional methods are mainly based on the knowledge of malicious URLs. Blacklist based and rule-based approaches use these known malicious URLs to detect phishing attacks. With the increase in the large number of phishing websites, it's very difficult to update the malicious URL database [3]. Hence, instead of traditional methods, machine learning approaches are used to identify suspicious websites.

Existing approaches based on machine learning use the features from URLs and websites to detect phishing attacks. Although these solutions achieve high accuracy, extracting features from the website content is time-consuming and has a high overhead. This issue is considered in most research works; however, a solution based on hyperlink features has not been extensively explored in existing works. Hyperlink features are also extracted from the website, but they are based on the links rather than the entire content available on the website; hence, they are much simpler to extract compared to content-based and network-based features.

This paper focuses on the features of URLs and hyperlinks used for detecting phishing websites. To achieve this, machine learning models such as SVM, Random Forest, XGBoost, and Stacking Classifier will be trained, and their performance will be evaluated. The performance is evaluated using Accuracy, Precision, Recall, and F1 score. The primary focus is to compare the accuracy and effectiveness of a machine learning model trained on URL features alone and on URL features combined with hyperlink features, to gain insight into whether hyperlink-based features contribute to improved performance.

The damage caused by compromising information systems is complex, including not only financial losses, but also damage to reputation, as well as litigation [4–6]. This multifaceted threat determines the critical importance of information security issues in the modern world.

2 Related Works

Singh and Charu [7] explained the approach by which machine learning techniques are used to identify phishing websites that closely resemble genuine websites. This work categorizes phishing websites into two primary types: social engineering attacks and malware attacks. This work also explains the most commonly used machine learning models. The effectiveness of these models in terms of detecting phishing websites is also discussed. The importance of identifying the most prominent features to detect phishing websites is also addressed. The performance of traditional phishing detection techniques and machine learning-based techniques is compared, and it is concluded that the machine learning approaches perform better in identifying phishing websites.

Akour et al. [8] utilized features based on URLs to detect phishing websites using various machine learning models. Its focus is mainly on the URLs; the content of the websites and network information are not considered in this work. Machine learning models, such as SVM, KNN, Logistic regression, and Naive Bayes, are used in the dataset with 147,000 URLs. From the URL dataset, around 111 different features are extracted,

mainly considering length, structure, and domain-related information. Experiments show that SVM performs better than other models, with an accuracy of 96.3%. This work demonstrates that phishing websites can be detected efficiently by simply using the URL.

Kumar et al. [9] used URL and domain-based features to detect phishing websites using different machine learning models. The significance of these features is also emphasized in this work. Results show that Naive Bayes performs better than other machine learning models with 98% accuracy. Alam et al. [10] used URL, HTML, and network-based information to detect phishing websites. Due to the integration of all these features, accuracy is increased. Decision trees and random forests are used for prediction. The accuracy is approximately 97%, indicating that the model predicts more accurately with hybrid features.

Lokesh et al. [11] extracted various features, including URLs and metadata based on HTML and JavaScript features, and used these to predict phishing website detection. Different machine learning models are employed in the proposed work, and the results demonstrate that random forests perform better, with low false positives and high accuracy. Rashid et al. [12] proposed a phishing website detection method using SVM. Web content and URL features are integrated to train the model. The UCI phishing dataset is used for training the model. The proposed model achieves an accuracy of approximately 95.66%, underscoring the importance of integrating various features.

Desai et al. [13] employ various preprocessing techniques to identify the most prominent features in the dataset. Features based on URL, domain, and website are identified as significant and used to train the models. An approach to integrate this proposed work with browsers is also suggested for detecting phishing websites. Although it integrates features of different types, it has a high rate of false positives, which is its main drawback. Saha et al. [14] proposed a multilayer perceptron to detect phishing websites. A Kaggle dataset containing information about 10,000 websites is used for training. Ten different features, including URL-based features and metadata, are used for training the model. It achieves an accuracy of about 93%. However, the proposed work has a high computational overhead due to the integration of multiple features.

Basit et al. [15] proposed an ensemble approach to detect phishing websites. An ensemble of ANN, KNN, Decision trees, and random forests is proposed in this work. Although various solutions [16–19] have been proposed for detecting phishing websites, integration of URLs with hyperlink features has not been explored much. Hence, in this proposed work, URL and hyperlink features are utilized to enhance the model's performance in detecting phishing websites. The proposed methodology is explained in detail in the next section.

3 Proposed Work

The proposed work follows a structured approach that includes data collection, feature processing, feature extraction, feature scaling, data splitting, model training, and performance evaluation. Machine learning models are trained with two feature sets: URL features and URL + Hyperlink features. By comparing these two feature sets, this work aims to provide valuable insights into the role of feature selection in improving phishing

detection models. The flow diagram of the proposed work is given in Fig. 1. Each phase is explained briefly in this section.

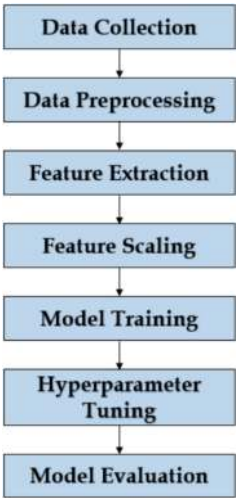


Fig. 1. Flowchart of the phishing detection system.

3.1 Data Collection

The first step in the implementation workflow is to collect the data to train the model. The dataset was collected from the Kaggle Website. The dataset contains labelled phishing and legitimate website URLs. The dataset was already balanced with both phishing and non-phishing website classes, so the model will not be biased. The dataset originally contained 89 features and 11,430 rows. However, only the URL and hyperlink features were selected for this work to maintain computational efficiency.

3.2 Data Preprocessing

Before a Machine learning model is trained, the data has to be pre-processed. The first step in preprocessing was to check for missing values and to maintain dataset integrity. After missing values were checked, duplicate URLs were searched and eliminated to prevent bias in model training. In the dataset, the dependent feature or the target feature, which was "status," was in a string format. The status column was encoded into numerical values, such as legitimate (0) and phishing (1).

3.3 Feature Extraction

Two different feature sets were extracted from the dataset to compare their effect on phishing detection. They are URL-only features and URL + hyperlink features.

- **URL-only features (44)**: These features were directly extracted from the lexical properties and structures of the URL, enabling phishing detection based solely on the URL. A total of 44 features were extracted. Some of the key extracted features are URL Length, IP, presence of special characters, Digit count in URL, and HTTP in URL.
- **URL + hyperlink features (54)**: This feature set includes all URL-only features along with hyperlink features extracted from webpage link structures. This feature set included 44 features of URL and 10 features of hyperlink. Some of the hyperlink features include the Number of hyperlinks, the ratio of internal links, the ratio of external links, the safe anchor, and the ratio of null hyperlinks. These two features set allow us to evaluate the impact of hyperlink-based features.

3.4 Feature Scaling and Data Splitting

After feature extraction, the data were split into an 80% training set and a 20% testing set, and numerical values were normalized where necessary. Stratified sampling was used to maintain a 50–50 class balance in both the training and testing sets. After the data were split, StandardScaler was applied to normalize numerical features. Scaling was performed separately on each feature set to maintain consistency. The data is now ready to be used by machine learning models for training.

3.5 Model Training

After the data was split and scaled, it was passed to the model for training. Machine learning models were trained on both feature sets separately to assess their effectiveness. The models used were Support Vector Machine (SVM), Random Forest, XGBoost, and Stacking classifier. These models are trained twice, using both URL features and URL + hyperlink features. Hyperparameter tuning is also used to optimize the performance of these models. RandomizedSearchCV is used for tuning the random forest model, and GridSearchCV is used for tuning SVM and XGBoost. Stacking classifier uses the fine-tuned models.

3.6 Model Evaluation

After training the models, they were tested and evaluated using the performance metrics. The models trained with URL and URL + Hyperlink features are tested, and performance is evaluated using accuracy, precision, recall, and F1-score. Comparison analysis is also conducted, as explained in the next section.

4 Performance Analysis

4.1 Performance of Models Using Only URL Features

The first assessment was performed using only URL only features. The result was obtained and is represented below on the Table 1. The performance shows that URL based features alone can provides a reasonable accuracy for detecting phishing websites. All the models achieved approximately 90% of accuracy. Out of all the models,

Stacking Classifier outperformed all the other models with an accuracy of 90.24% which demonstrated the effectiveness of ensemble model in classification performance. Random Forest with 90.07% and XGBoost with 89.90% performed slightly better than Support Vector Machine Model which suggests tree-based models are capable of capturing URL patterns and characteristics.

Table 1. Performance of models using only URL features.

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	88.28	88.14	88.45	88.30
XGBoost	89.90	89.65	90.20	89.93
Random Forest	90.07	89.35	90.99	90.16
Stacking Classifier	90.24	90.00	90.55	90.27

The SVM performed the worst among all the other models, with an accuracy of 88.28% which indicates that URL-based feature alone may not always provide clear feature separability for optimal classification. While URL-based features provide decent accuracy, they have certain limitations. These features focus on the lexical properties of URLs, which fail to provide additional information about external redirection, hidden links, and other web page behaviors that could further enhance the accuracy of the phishing detection system.

4.2 Integration of URL and Hyperlink Features

The URL and hyperlink features are integrated to train the models, and the performance is analysed in this section. Hyperlink features include information regarding the structure of the websites. The performance analysis of the models using these integrated features is tabulated in Table 2. As shown, the models perform better due to the inclusion of hyperlink features. The stacking classifier achieves a high accuracy of 93.70% due to the integration of multiple models. XGBoost and Random Forest achieved accuracies of 93.22% and 93.13% respectively, which is higher compared to the performance with URL features excluding hyperlink features.

Table 2. Performance of models using URL + hyperlink features.

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	91.29	91.48	91.08	88.30
XGBoost	93.13	93.55	92.65	89.93
Random Forest	93.22	94.11	92.11	90.16
Stacking Classifer	93.70	94.01	90.55	90.27

The Support Vector Machine model also showed significant improvement, with an accuracy of 91.29%, but it is still less than that of the other models. This suggests that hyperlink-based features are more beneficial for ensemble and tree-based models than for Support Vector Machine, which relies on hyperplane classification. The performance is also graphically illustrated in Fig. 2. This graph clearly illustrates that all the models showed improved accuracy when trained with URL + hyperlink features. The results highlight that adding hyperlink features improves model accuracy and enables the models to distinguish between legitimate and phishing websites better by capturing additional behaviors and patterns.

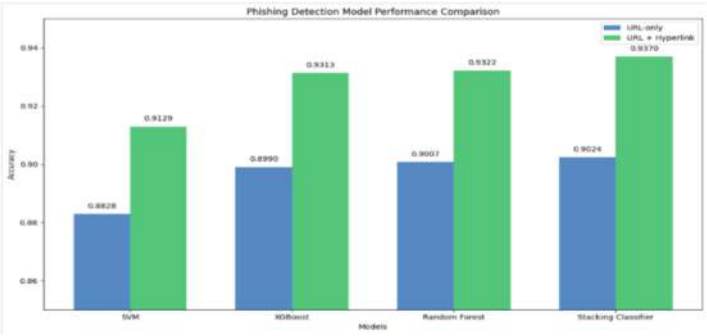


Fig. 2. Performance analysis—accuracy.

Through this analysis, it is identified that hyperlink-based features are significant in accurately identifying phishing websites. Rather than relying entirely on expensive content-based and network-based features, hyperlink features can be integrated to train the models, which provides a cost-effective solution for detecting phishing websites. These features work well with tree-based models, which also improves the performance of the proposed ensemble approach.

5 Conclusion

In this paper, the significance of hyperlink features for the detection of phishing websites is highlighted by training different ML models using hyperlink features in addition to URL features. Since dealing with hyperlink features is easier and less expensive compared to network-based and content-based features on websites, the proposed work provides a cost-effective solution to identify phishing websites. URL and hyperlink features are integrated to train the model. Among the models trained, XGBoost and random forest perform better, and the stacking classifier with these models yields an accuracy of 93.70% which is high compared to other solutions used. Hence, results emphasize that hyperlink features have a major impact on identifying phishing websites.

References

1. Anti-Phishing Working Group (APWG): Phishing Activity Trends Report Q1 2025. https://apwg.org/trendreports
2. Marchal, S., François, J., State, R., Engel, T.: PhishStorm: detecting phishing with streaming analytics. IEEE Trans. Netw. Serv. Manag. **11**(4), 458–471 (2014)
3. Verma, R., Das, A.: What's in a URL: fast feature extraction and malicious URL detection. In: IWSPA '17: Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics, pp. 55–63. https://doi.org/10.1145/3041008.3041016
4. Parfenov, D., Bolodurina, I., Lapina, M.: Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Conf. Ser. Mater. Sci. Eng. **873**, 012020 (2020). https://doi.org/10.1088/1757-899X/873/1/012020
5. Maksimova, E.A., Lapina, M.A., Lapin, V.G., Rusakov, A.M.: Anthropomorphic model of states of subjects of critical information infrastructure under destructive influences. In: Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I. (eds.) Mathematics and its Applications in New Computer Systems. MANCS 2021. Lecture Notes in Networks and Systems, vol. 424. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97020-8_51
6. Lapina, M., Sajid, M., Bagautdinova, A., Lapin, V., Tokmakova, M.: Comparative characteristics of the methods of using digital watermarks to protect text information. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds.) AISMA-2023: International Workshop on Advanced Information Security Management and Applications. AISMA 2023. Lecture Notes in Networks and Systems, vol. 1207. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77229-0_15
7. Singh, C.: Phishing website detection based on machine learning: a survey. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE (2020)
8. Akour, I., et al.: Using classical machine learning for phishing websites detection from URLs. J. Manag. Inf. Decis. Sci. **24**(6), 1–15 (2021)
9. Kumar, J., et al.: Phishing website classification and detection using machine learning. In: 2020 International Conference on Computer Communication and Informatics (ICCCI). IEEE (2020)
10. Alam, M.N., et al.: Phishing attacks detection using machine learning approach. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE (2020)
11. Lokesh, H., Gururaj, BoreGowda, G.: Phishing website detection based on effective machine learning approach. J. Cyber Secur. Technol. **5**(1), 1–14 (2021)
12. Rashid, J., et al.: Phishing detection using machine learning technique. In: 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH). IEEE (2020)
13. Desai, P., Shah, M.: Phishing website detection using machine learning: a comprehensive study. Int. J. Multidiscip. Res. (2023)
14. Saha, I., et al.: Phishing attacks detection using deep learning approach. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE (2020)
15. Basit, A., et al.: A novel ensemble machine learning method to detect phishing attack. In: 2020 IEEE 23rd International Multitopic Conference (INMIC). IEEE (2020). Dharani, M., et al.: Detection of phishing websites using ensemble machine learning approach. In: ITM Web of Conferences, vol. 40. EDP Sciences (2021)
16. Gontla, B.K., et al.: A machine learning approach to identify phishing websites: a comparative study of classification models and ensemble learning techniques. EAI Endorsed Trans. Scalable Inf. Syst. **10**(5) (2023)

17. Innab, N., et al.: Phishing attacks detection using ensemble machine learning algorithms. Comput. Mater. Contin. **80**(1) (2024)
18. Adeyemo, V.E., et al.: Ensemble-based logistic model trees for website phishing detection. In: Advances in Cyber Security: Second International Conference, ACeS 2020, Penang, Malaysia, December 8–9, 2020, Revised Selected Papers 2. Springer Singapore (2021)
19. Ajayi, Ojo, Adetunmbi, Adebayo, Olowookere, Toluwase, Sodiya, Simon (2023) Phishing Detection: Performance Evaluation of both Ensemble and Classical Machine Learning Models.

Comprehensive Assessment of Ethanol and Extract Content in Wine Products

Anatoliy Kazak[1](✉), Ruslan Timofeev[2], Nikolay Oleinikov[1], Marina Rudenko[1], Anisa Atik[1], and Anastasia Bobarykina[1]

[1] V.I. Vernadsky Crimean Federal University, 295007 Simferopol, Russia
kazak_a@mail.ru

[2] "All-Russian National Research Institute of Viticulture and Winemaking" Magarach, "RAS", 298600 Yalta, Russia

Abstract. This study focuses on developing a methodological approach to rapidly determining the mass concentration of extractives and the volume fraction of ethanol in wine distillates based on the optical characteristics of the solutions. The study analyzed existing technical and chemical methods for determining the mass concentration of extractives and ethanol during wine production. The relationship between the optical density and refractometric quantitative values of sugars and the ethanol concentration in ethanol-water reaction solutions of wine origin was established. The influence of sugars on the densimetric and refractometric parameters in grape-based products was also determined. The results of the study of the complex influence of sugar and ethanol content on the optical properties of ethanol-water reaction solutions of viticultural and winemaking nature are valuable in the creation of equipment that allows for the rapid assessment of the proportion of each individual component in the matrix of the chemical composition and organoleptic properties of grape wines and aged high-proof alcoholic beverages, as well as the ability to make an objective decision on the quality of finished wine products.

Keywords: Optical properties · Rapid methods · Instrumental methods of analysis · Extractive substances · Product quality

1 Introduction

Wine researchers have always been challenged by the need to quickly and accurately determine the quality parameters of biochemical solutions, including grape processing products and wine distillates obtained by distilling wine products. The fastest and easiest express methods to implement are those based on measuring and evaluating the optical properties of reaction mixtures consisting primarily of extractive substances, ethanol, and carbohydrates.

The generally accepted method to date for determining the mass concentration of total extract in wine products with an ethanol volume fraction of 35.0% or higher is the OIV-MA-AS2-03B method [1–4]. The working range within which the mass concentration of total extract can be determined with a high degree of accuracy is from 0.1 to 25.0 g/dm^3.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 205–214, 2027.
https://doi.org/10.1007/978-3-032-27834-0_20

The generally accepted and approved method in the world, which allows for a highly accurate determination of the concentration of ethyl alcohol, is the method of the International Organization of Vine and Wine, the essence of which boils down to the fact that the test sample is subjected to distillation, and then the resulting solution is analyzed to determine the number of volume units of pure alcohol, calculated per 100 volume units of the product, at a temperature of 20 °C. The OIV-MA-AS312-01A method is based on measuring the density of the distillate and does not take into account the influence of extractive substances previously removed during the distillation process [4, 5].

Methods for determining total extract in wine products and the volume fraction of ethanol in distillates are characterized by accuracy and high labor intensity associated with the additional stage of destruction of the primary composition of the test object, which specialists in various independent laboratories encounter when analyzing the same samples. Therefore, the existing limitations prevent these methods from being recommended as rapid, rapid monitoring tools in relevant winemaking processes [6–8].

Researchers are currently faced with the task of developing a fast and reliable method for the simultaneous identification of both alcohols and compounds in a test sample of alcoholic beverages, the total content of which forms the total and reduced extracts, based on the introduction of modern methods of chemical analytical analysis, such as gas chromatography (GC), high-performance liquid chromatography (HPLC), IR spectroscopy and others.

The modified FTIR spectroscopy method has the greatest practical potential for developing such a rapid method [9–12]. The FTIR spectroscopy method is implemented in the Alex 500 analyzer manufactured by the Austrian company Anton Paar GmbH. The operating principle of the analyzer is that the volume fraction of ethyl alcohol is determined spectrometrically in the IR region, and the value of the total extract is determined as the result of an algebraic operation to establish the difference in density values.

During the analysis carried out in accredited laboratories and the protocols drawn up in them, it was established that the Alex 500 analyzer has a significant drawback, which is that the values obtained with it help do not fully correspond to the data obtained using the arbitration analysis methods OIV-MA-AS312-01A and OIV-MA-AS2-03B. This discrepancy in the obtained results is due to differences in the physical principles underlying the use of Alex 500 and arbitration methods of analysis.

Refractodensimetry is a promising and in-demand physicochemical method that allows the development of a rapid method for determining alcohol content and the total content of non-volatile substances without the use of destructive procedures on the analyzed sample.

2 Materials and Methods

The test subjects were complex composite solutions containing varying base components found in wine products (prepared water with a resistivity in the range of 0.6–0.8 MΩ, chemically pure β-D-fructofuranosyl α-D-glucopyranoside, and rectified ethyl hydrate with a concentration of at least 96.0% vol.). The concentration of ethyl hydrate was determined by the method OIV-MA-AS312 01A; the concentration of non-volatile substances was determined by the method OIV-MA-AS2-03B [13].

Mathematical processing of the experimental data was performed using a licensed Microsoft application package. A schematic diagram and sequence of analyses are presented in Fig. 1.

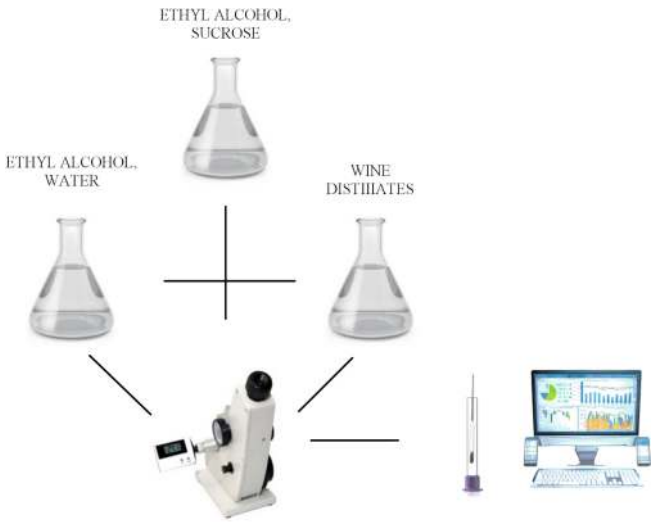


Fig. 1. A schematic diagram and sequence of analyses.

Optical properties and their changes during analysis were determined using an IRF-23 refractometer as the concentration of rectified ethyl hydrate in typical solutions increased, with the maximum value being taken for a ratio of 70 parts ethyl hydrate to 30 parts treated water. The working length of the sample for identification was empirically established at 589 nm. The next step was the preparation of standard solutions of β-D-fructofuranosyl α-D-glucopyranoside with a multiple ratio from the blank to a solution with a ratio of 75 parts of prepared water and 25 parts of β-D-fructofuranosyl α-D-glucopyranoside. The values determined during the analyses were compared using the function approximation method using polynomials [14–16].

3 Results and Discussion

The readings obtained during the measurement process were passed through a system with an interpolating piecewise cubic polynomial with specified tangents at the endpoints (at the boundaries of the curve sections) and are presented in Table 1.

The Fig. 2 presents experimental data confirming the relationship between the refractometer's digital carbohydrate scale readings and the concentration of ethyl hydrate in typical solutions.

The Fig. 3 presents data expressing the numerical and physical dependence of the hydrometer scale and the quantitative characteristics of ethyl hydrate in the presence of a component such as β-D-fructofuranosyl α-D-glucopyranoside.

Table 1. Values of refractometer and alcohol meter readings for various ethanol solutions.

% about.	n	IN	% about.	n	IN	% about.	n	IN
0.0	1.3330	0.00	32.0	1.3513	12.26	64.0	1.3630	19.52
1.0	1.3335	0.34	33.0	1.3519	12.61	65.0	1.3632	19.64
2.0	1.3339	0.68	34.0	1.3524	12.95	66.0	1.3634	19.77
3.0	1.3344	1.02	35.0	1.3529	13.25	67.0	1.3636	19.90
4.0	1.3349	1.32	36.0	1.3534	13.55	68.0	1.3638	20.02
5.0	1.3354	1.64	37.0	1.3539	13.85	69.0	1.3640	20.12
6.0	1.3359	2.01	38.0	1.3543	14.15	70.0	1.3642	20.21
7.0	1.3364	2.38	39.0	1.3548	14.46	71.0	1.3643	20.31
8.0	1.3370	2.77	40.0	1.3553	14.76	72.0	1.3645	20.40
9.0	1.3375	3.15	41.0	1.3558	15.06	73.0	1.3646	20.50
10.0	1.3381	3.51	42.0	1.3562	15.31	74.0	1.3648	20.59
11.0	1.3387	3.91	43.0	1.3566	15.56	75.0	1.3650	20.69
12.0	1.3392	4.32	44.0	1.3570	15.82	76.0	1.3651	20.76
13.0	1.3398	4.72	45.0	1.3574	16.07	77.0	1.3652	20.83
14.0	1.3405	5.11	46.0	1.3578	16.31	78.0	1.3653	20.89
15.0	1.3411	5.55	47.0	1.3582	16.55	79.0	1.3654	20.95
16.0	1.3416	5.90	48.0	1.3585	16.75	80.0	1.3655	20.98
17.0	1.3423	6.36	49.0	1.3588	16.95	81.0	1.3656	21.04
18.0	1.3429	6.76	50.0	1.3592	17.17	82.0	1.3657	21.11
19.0	1.3435	7.17	51.0	1.3595	17.41	83.0	1.3657	21.12
20.0	1.3442	7.60	52.0	1.3599	17.62	84.0	1.3657	21.12
21.0	1.3448	7.99	53.0	1.3602	17.80	85.0	1.3658	21.16
22.0	1.3454	8.43	54.0	1.3605	17.98	86.0	1.3658	21.16
23.0	1.3460	8.81	55.0	1.3608	18.15	87.0	1.3657	21.12
24.0	1.3466	9.18	56.0	1.3610	18.32	88.0	1.3656	21.08
25.0	1.3472	9.56	57.0	1.3613	18.49	89.0	1.3656	21.08
26.0	1.3478	10.01	58.0	1.3616	18.66	90.0	1.3655	21.01
27.0	1.3484	10.41	59.0	1.3619	18.81	91.0	1.3654	20.94
28.0	1.3490	10.77	60.0	1.3621	18.96	92.0	1.3653	20.86
29.0	1.3496	11.16	61.0	1.3624	19.11	93.0	1.3650	20.74
30.0	1.3502	11.53	62.0	1.3626	19.26	94.0	1.3648	20.59
31.0	1.3507	11.90	63.0	1.3628	19.39	95.0	1.3645	20.42

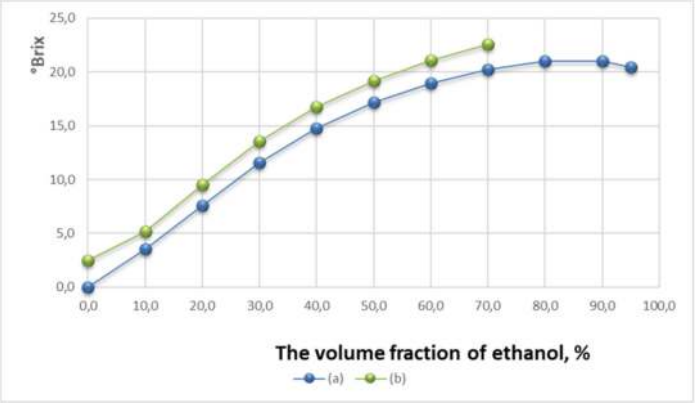


Fig. 2. Refractometric values taking into account the quantitative characteristics of ethyl hydrate **a**–solution of ethyl hydrate in water; **b**–solution of β-D-fructofuranosyl α-D-glucopyranoside in the water-ethyl hydrate system.

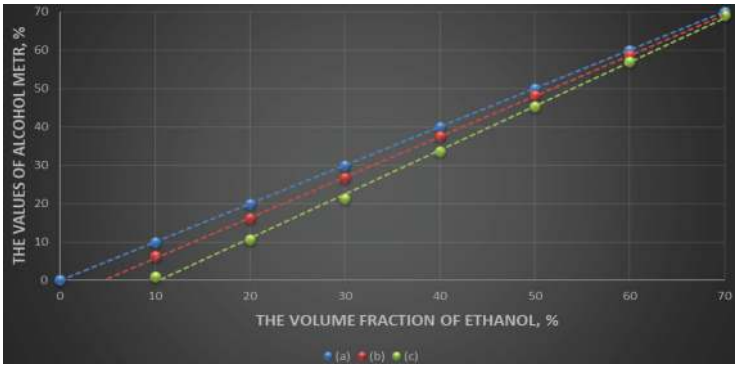


Fig. 3. Correlation of alcohol meter values and ethyl alcohol concentration: **a**–solution of ethyl hydrate in water; **b**–solution of β-D-fructofuranosyl α-D-glucopyranoside in the water-ethyl hydrate system with mass concentration of 10 g/dm^3; **c**–solution of β-D-fructofuranosyl α-D-glucopyranoside in the water-ethyl hydrate system with mass concentration of 25 g/dm^3.

Based on the obtained data (Fig. 3), a linear relationship was established between the hydrometer readings and the quantitative characteristics of the ethyl hydrate solution containing β-D-fructofuranosyl α-D-glucopyranoside.

Processing the values of optical measurements of refractive indices and densimetric characteristics of typical solutions with previously known quantitative characteristics of ethyl hydrate and β-D-fructofuranosyl α-D-glucopyranoside made it possible to obtain numerical values reflected in the dependence of the concentration of ethyl hydrate (Table 2) and the mass concentration of non-volatile substances (Table 3) on the numerical values of the hydrometer and the readings of the carbohydrate value of the refractometer.

Table 2. The value of ethanol concentration taking into account the readings of the refractometer and alcohol meter.

α	The values of refractometer scale, %										
	13	14	15	16	17	18	19	20	21	22	23
26	30.57	32.15	33.74	35.34	–	–	–	–	–	–	–
27	30.98	32.56	34.15	35.74	–	–	–	–	–	–	–
28	31.40	32.98	34.57	36.15	–	–	–	–	–	–	–
29	31.83	33.41	35.00	36.58	–	–	–	–	–	–	–
30	32.26	33.85	35.43	37.01	38.60	–	–	–	–	–	–
31	32.71	34.29	35.87	37.45	39.03	–	–	–	–	–	–
32	33.16	34.74	36.32	37.89	39.48	–	–	–	–	–	–
33	33.63	35.21	36.78	38.36	39.94	–	–	–	–	–	–
34	34.11	35.68	37.25	38.83	40,40	41.99	–	–	–	–	–
35	34.60	36.17	37.75	39.32	40.89	42.49	–	–	–	–	–
36	35.11	36.68	38.25	39.82	41.39	42.99	–	–	–	–	–
37	35.63	37.19	38.76	40.33	41.92	43.50	–	–	–	–	–
38	36.15	37.72	39.29	40.85	42.45	44.02	–	–	–	–	–
39	–	38.26	39.83	41.40	43.00	44.55	46.18	–	–	–	–
40	–	38.80	40.37	41.96	43.54	45.08	46.72	–	–	–	–
41	–	39.36	40.91	42.52	44.09	45.63	47.27	–	–	–	–
42	–	–	41.50	43.10	44.65	46.20	47.84	–	–	–	–
43	–	–	42.11	43.69	45.24	46.80	48.42	–	–	–	–
44	–	–	42.72	44.29	45.84	47.40	49.02	50.44	–	–	–
45	–	–	43.34	44.89	46.47	48.02	49.61	51.01	–	–	–
46	–	–	–	45.52	47.10	48.65	50.21	51.58	–	–	–
47	–	–	–	46.52	48.09	49.62	51.13	52.51	–	–	–
48	–	–	–	46.83	48.40	49.92	51.41	52.80	–	–	–
49	–	–	–	47.51	49.08	50.55	52.04	53.44	54.80	–	–
50	–	–	–	48.20	49.75	51.20	52.71	54.08	55.40	–	–
51	–	–	–	–	50.42	51.86	53.39	54.74	56.02	–	–
52	–	–	–	–	51.10	52.57	54.08	55.40	56.65	–	–
53	–	–	–	–	51.80	53.29	54.78	56.06	57.29	–	–
54	–	–	–	–	52.54	54.02	55.49	56.75	57.93	–	–
55	–	–	–	–	53.31	54.77	56.21	57.44	58.59	–	–

(*continued*)

Table 2. (*continued*)

α	The values of refractometer scale, %										
	13	14	15	16	17	18	19	20	21	22	23
56	–	–	–	–	–	55.52	56.94	58.14	59.26	60.18	–
57	–	–	–	–	–	56.31	57.70	58.87	59.96	60.83	–
58	–	–	–	–	–	57.10	58.47	59.61	60.66	61.48	–
59	–	–	–	–	–	57.91	59.26	60.37	61.37	62.14	–
60	–	–	–	–	–	58.74	60.06	61.15	62.10	62.81	–
61	–	–	–	–	–	59.60	60.89	61.94	62.84	63.49	–
62	–	–	–	–	–	–	61.74	62.75	63.60	64.18	–
63	–	–	–	–	–	–	62.61	63.58	64.37	64.89	–
64	–	–	–	–	–	–	63.50	64.43	65.16	65.60	65.72
65	–	–	–	–	–	–	64.42	65.29	65.96	66.33	66.36
66	–	–	–	–	–	–	65.35	66.18	66.78	67.07	67.02
67	–	–	–	–	–	–	66.31	67.08	67.61	67.82	67.69
68	–	–	–	–	–	–	67.29	68.00	68.46	68.59	68.36
69	–	–	–	–	–	–	–	68.95	69.33	69.38	69.05
70	–	–	–	–	–	–	–	69.92	70.23	70.18	69.73

The use of a hydrometer to measure densimetric characteristics ensures the accuracy of determining the density of liquid systems, and the direct use of hydrometer readings allows one to obtain final results that do not require additional mathematical calculations.

The numerical rank of the tables is given in accordance with the reflection of the accuracy of the calculation of the quantitative characteristics of ethyl hydrate and non-volatile substances.

4 Conclusions

The study allowed us to obtain numerical relationships between physicochemical parameters such as the refractive index of ethyl hydrate-based solutions and treated water with a resistivity of 0.6–0.8 MΩ and the refractometer's carbohydrate scale readings for determining the quantitative characteristics of ethyl hydrate, expressed as a percentage by volume. A methodology was developed for the accelerated determination of the quantitative characteristics of ethyl hydrate and non-volatile substances in wine products, based on the use of a refractometer equipped with a carbohydrate scale and a hydrometer.

Table 3. Quantitative characteristics of non-volatile substances taking into account the readings of the refractometer and hydrometer (at 20 °C).

A	The values of refractometer scale, %										
	13	14	15	16	17	18	19	20	21	22	23
26	14.90	20.52	26.19	32,00	–	–	–	–	–	–	–
27	13.24	18.87	24.55	30.52	–	–	–	–	–	–	–
28	11.56	17.20	22.90	29.03	–	–	–	–	–	–	–
29	9.82	15.47	21.18	27.48	–	–	–	–	–	–	–
30	8.06	13.72	19.57	25.93	32.39	–	–	–	–	–	–
31	6.27	11.95	17.97	24.35	30.84	–	–	–	–	–	–
32	4.45	10.13	16.33	22.74	29.26	–	–	–	–	–	–
33	2.53	8.30	14.63	21.06	27.61	–	–	–	–	–	–
34	0.60	6.54	12.90	19.37	25.96	32.96	–	–	–	–	–
35	−1.43	4.72	11.10	17.60	24.26	31.43	–	–	–	–	–
36	−3.44	2.86	9.28	15.80	22.60	29.90	–	–	–	–	–
37	−5.37	0.96	7.41	13.98	20.96	28.32	–	–	–	–	–
38	−7.35	−0.98	5.49	12.14	19.31	26.71	–	–	–	–	–
39	–	−2.97	3.54	10.35	17.64	25.07	33.29	–	–	–	–
40	–	−4.96	1.61	8.60	15.95	23.42	31.79	–	–	–	–
41	–	−7.00	−0.35	6.85	14.23	21.83	30.28	–	–	–	–
42	–	–	−2.22	5.07	12.49	20,24	28.74	–	–	–	–
43	–	–	−4.13	3.21	10.71	18.59	27.15	–	–	–	–
44	–	–	−6.03	1.35	9.00	16.94	25.57	33.81	–	–	–
45	–	–	−7.96	−0.54	7.27	15.28	23.96	32.21	–	–	–
46	–	–	–	−2.36	5.54	13.60	22.33	30.60	–	–	–
47	–	–	–	−5.13	2.87	11.02	19.77	28.28	–	–	–
48	–	–	–	−5.98	2.04	10.22	18.98	27.59	–	–	–
49	–	–	–	−7.81	0.28	8.47	17:30	26.10	35.56	–	–
50	–	–	–	−9.65	−1.50	6.71	15.77	24.61	34.11	–	–
51	–	–	–	–	−3.31	4.94	14.23	23.12	32.66	–	–
52	–	–	–	–	−5.16	3.30	12.67	21.60	31.20	–	–
53	–	–	–	–	−7.00	1.70	11.12	20.09	29.75	–	–
54	–	–	–	–	−8.72	0.08	9.56	18.59	28.31	–	–
55	–	–	–	–	−10.38	−1.53	8.00	17.09	26.88	–	–

(*continued*)

Table 3. (*continued*)

A	The values of refractometer scale, %										
	13	14	15	16	17	18	19	20	21	22	23
56	–	–	–	–	–	−3.12	6.47	15.63	25.47	35.47	–
57	–	–	–	–	–	−4.75	4.91	14.13	24.08	34.12	–
58	–	–	–	–	–	−6.35	3.39	12.67	22.71	32.80	–
59	–	–	–	–	–	−7.93	1.88	11.27	21.37	31.50	–
60	–	–	–	–	–	−9.48	0.41	9.90	20.04	30.21	–
61	–	–	–	–	–	−11.02	−1.01	8.54	18.74	28.93	–
62	–	–	–	–	–	–	−2.41	7.20	17.45	27.66	–
63	–	–	–	–	–	–	−3.77	5.90	16.18	26.40	–
64	–	–	–	–	–	–	−5.10	4.63	14.93	25.17	35.29
65	–	–	–	–	–	–	−6.39	3.38	13.71	23.97	34.08
66	–	–	–	–	–	–	−7.64	2.16	12.52	22.79	32.88
67	–	–	–	–	–	–	−8.87	0.97	11.37	21.64	31.67
68	–	–	–	–	–	–	−10.05	−0.17	10.26	20.51	30.46
69	–	–	–	–	–	–	–	−1.27	9.17	19.38	29.26
70	–	–	–	–	–	–	–	−2.32	8.11	18.26	28.06

Conflicts of Interest. The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding. The authors acknowledge the financial support provided by the Russian Science Foundation, grant No. 24-25-20001.

Data Availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Noiseux, I., Long, W., Cournoyer, A., Vernon, M.: Simple fiber-optic-based sensors for process monitoring: an application in wine quality control monitoring. Appl. Spectrosc. **58**(8), 1010–1019 (2004). https://doi.org/10.1366/0003702041655476
2. Cai, C. et al.: Comparison of methods for predicting the compositional dependence of the density and refractive index of organic-aqueous aerosols. J. Phys. Chem. A. **120**(33), 6604–6617 (2016). https://doi.org/10.1021/acs.jpca.6b05986
3. Plugatar, Y. et al.: Prediction of ethanol content and total extract using densimetry and refractometry. Beverages. **9**(2), 31 (2023). https://doi.org/10.3390/beverages9020031
4. OIV: Compendium of International Methods of Wine and Must Analysis; International Organization of Vine and Wine: Dijon, France (2019)

5. Iland, P., Ewart, A., Sitters, J., Markides, A., Bruer, N.: Techniques for Accurate Chemical Analysis and Quality Monitoring during Winemaking. Campbell Town, Australia, Patrick Iland Promotions (2000)
6. Timofeev, R.: Refractodensimetric method for determining the volume fraction of ethyl alcohol in wines and winy beverages. Proc. Voronezh State Univ. Eng. Technol. **82**(3), 104–109 (2020). https://doi.org/10.20914/2310-1202-2020-3-104-109
7. Debebe, A., Redi-Abshiro, M., Chandravanshi, B.S.: Nondestructive determination of ethanol levels in fermented alcoholic beverages using fourier transform midinfrared spectroscopy. Chem. Central J. **11**(1), 27 (2017). https://doi.org/10.1186/s13065-017-0257-5
8. Timofeev, R.G.: Development of a refractodensimetric method for determining the content of ethyl alcohol and total wine extract by a computer. Sci. J. NRU ITMO Process. Food Prod. Equip. **4**(46), 3–11 (2020). https://doi.org/10.17586/2310-1164-2020-10-4-3-11
9. Thanasi, V., Caldeira, I., Santos, L., Ricardo-da-Silva, J., Catarino, S.: Simultaneous determination of ethanol and methanol in wines using FTIR and PLS regression. Foods. **13**(18), 2975 (2024). https://doi.org/10.3390/foods13182975
10. Kolesnov, A. et al.: Correct determination of alcoholic strength in alcoholic products. BIO Web Conf. **68**, 02030 (2023). https://doi.org/10.1051/bioconf/20236802030
11. Johnson, J.B., Timofeev, R., Kazak, A., Grishin, Y., Solovyova, L., Rudenko, M.: A study of the UV spectral features in wine and their correlation with phenolic constituents. Front. Biosci. (Elite Ed.). **16**(2), 16 (2024). https://doi.org/10.31083/j.fbe1602016
12. Paolini, M., Tonidandel, L., Larcher, R.: Development, validation and application of a fast GC-FID method for the analysis of volatile compounds in spirit drinks and wine. Food Control. **136**(2), 108873 (2022). https://doi.org/10.1016/j.foodcont.2022.108873
13. Kazak, A. et al.: Application of grape stems extract to increase the antioxidant capacity of whiskey. Int. J. Food Sci. **2024**, 71909030 (2024). https://doi.org/10.1155/2024/7199030
14. Georgieva, I., Hofreither, C.: A Newton's method for best uniform polynomial approximation. Numer. Algorithms. **93**(4), 1741–1758 (2022). https://doi.org/10.1007/s11075-022-01487-5
15. Ali, I., Yaseen, M., Abdullah, M., Khan, S., Belgacem, F.B.M.: An innovative numerical method utilizing novel cubic B-spline approximations to solve burgers' equation. Mathematics. **11**, 4079 (2023). https://doi.org/10.3390/math11194079
16. Yousaf, M.Z. et al.: A novel quintic B-spline technique for numerical solutions of the fourth-order singular singularly-perturbed problems. Symmetry. **15**, 1929 (2023). https://doi.org/10.3390/sym15101929

Analysis of Methods for Detecting Anomalies in Information Systems Based on Containerization Technology

Egor Kenkov[1(✉)], Andrey Pavlov[1], and Mikhail Rusanov[2]

[1] North Caucasus Federal University, Stavropol 355017, Russia
kenkov2045@yandex.ru, Losde5530@gmail.com
[2] Moscow University of Finance and Law, Moscow 115191, Russia
mix.rusanoff@yandex.ru

Abstract. This study addresses the pressing issue of ensuring information security in information systems based on containerization technology. Its widespread adoption provides scalability and flexibility, but expands the attack surface due to the shared host operating system kernel. The objective of the study is a practical comparative analysis of machine learning-based anomaly detection methods in container environments using various data sources: system calls, application logs, and host resources. To address this problem, we deployed a test data collection and processing system as part of the experiment. We assessed the effectiveness of a combined approach, which integrates technologies effective for individual types of input data, comparing it to a universal solution analyzing all sources simultaneously. The results demonstrate that combining several methods improves the accuracy of anomaly classification through synergistic effects. However, this approach introduces a number of additional complexities to the development and integration of such a solution. While the universal solution demonstrates somewhat lower effectiveness, it has the potential to be adapted to the specific needs of a specific information system. Based on the obtained results, we formulated a direction for further work aimed at developing flexible and comprehensive solutions for monitoring the security of containerized applications.

Keywords: Information security · Artificial intelligence · Machine learning · Containerization · Anomaly detection

1 Introduction

1.1 Analysis of Information Security Issues in Information Systems Based on Containerization Technology

Virtualization tools are an integral part of modern information technology, providing cost-effectiveness, scalability, and flexibility for the infrastructure of enterprises and organizations of all sizes. Recently published studies question the security of systems based on virtualization tools [1–3]. These studies indicate that virtualization, in its broadest sense, is susceptible to threats and various attacks (escape from a container, side-channel attacks, privilege escalation).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 215–223, 2027.
https://doi.org/10.1007/978-3-032-27834-0_21

The most popular and widespread implementation of containerization technology is Docker. Unlike classic hypervisor-based virtualization, which emulates hardware and requires running a full-fledged guest operating system in each virtual machine, Docker implements virtualization at the operating system level [4]. Docker's architecture based on a client-server model, with the key component being the dockerd daemon, which manages the container lifecycle, images, networks, and data volumes via the Docker API [5].

The technological basis for process isolation and resource management in Docker are the Linux kernel mechanisms: namespaces and control groups (cgroups) [4]. Namespaces provide isolation of global system resources for a group of processes, creating the illusion that a container has its own instance of resources such as the process tree (PID), network subsystem (Net), and file system mount points (Mnt). This prevents processes from one container from interfering with the operation of other containers or the host system. The cgroups mechanism is responsible for accounting for and limiting resource consumption (CPU, memory, disk, network) for each container, thereby ensuring control over the resources used by containers and preventing a single container from exhausting the host machine's resources [6].

A Docker container is created from an immutable image, which is essentially a multi-layered filesystem containing the application with all its dependencies, libraries, and configuration. Due to its layered structure, a Docker image enables efficient storage and reuse of common components across different images. A Docker image built using a Dockerfile, a file that describes a declarative script with a sequence of commands [5]. Completed Docker images are stored and distributed through public or local image registries, such as the public Docker Hub or private repositories.

Therefore, Docker containers are characterized by extremely low system resource consumption, fast startup times (up to a few seconds), and efficient resource utilization compared to virtual machines [4, 6]. However, as discussed above and confirmed in research [5, 6], containerization technology has significant security issues, as the use of a shared host kernel creates a wider attack surface due to kernel vulnerabilities that can compromise all containers on the host. This is one of the key factors calling into question the security of containerization systems.

Russia's government recognizes the urgency of securing virtualization and containerization technologies as a priority. Presidential Decree No. 203 of May 9, 2017, "On the Strategy for the Development of the Information Society in the Russian Federation for 2017–2030," identifies key objectives as ensuring information security and replacing imported software with Russian equivalents. The security of virtualization and containerization platforms, as well as ensuring this security through Russian equivalents of foreign software, are critical to achieving these goals, as their vulnerabilities directly threaten the resilience of critical information infrastructure (CII) and data confidentiality.

1.2 Analysis of the Current State of the Research Topic

Recently, the use of machine learning methods for anomaly detection has become a popular trend [7–9], as these methods provide higher accuracy in anomaly detection. However, anomaly detection in containerized environments has received special attention in many studies [10, 11].

Consider possible solutions based on the use of machine learning methods that implement various input data: system calls, logs, host operating system resources (memory usage, processor utilization).

Solution [12] examines the use of system calls to detect anomalies in container environments. This study analyzes the effectiveness of four machine learning algorithms to identify the most promising tools for solving this problem. This study has a limitation in that it collects system calls exclusively from processes running inside a single container, rendering the system insufficiently horizontally scalable.

In contrast to the previous study, one may refer to a solution [13], which creates a graph model reflecting the dependencies between system calls within a container. This solution is capable of collecting information from all containers deployed on the host machine, but it is unable to find dependencies and relationships between system calls of multiple interacting containers, as it processes the system calls of each container separately, as in the previous study.

However, the solution [14] already considers a method for detecting anomalies not only individually for each container through system call analysis, but also applies to the analysis of system calls across a collection of containers that generate system calls when interacting with each other. For example, a container with a frontend application may constantly send requests to a container with a database, and this study takes into account the communication between these two containers.

The study [15] collects container system calls for further analysis using an autoencoder machine learning model. We can also apply this approach to analyze structured logs. The solution [16] describes a comparison of five popular deep network models for log analysis and anomaly detection, including an autoencoder. However, this immediately introduces the non-trivial task of processing logs for feeding them to the autoencoder, so the study [17] thoroughly examines and addresses the problem of log normalization and vectorization prior to subsequent analysis. However, since this autoencoder does not rely on an LSTM model, it effectively handles noisy log data but struggles when trained on unlabeled datasets.

A study [18] examines two methods for analyzing cache access by the host system running a containerization tool. Both methods have minimal anomaly detection time and low computational complexity, as they analyze a minimal number of metrics. Furthermore, the system collects and analyzes these metrics directly at the host level, rather than separately for individual containers or virtual machines. However, this solution also has its drawbacks, as it is impossible to localize the container in which the anomaly occurred, and it can only detect Memory DoS anomalies.

The paper [19] explores an approach to detecting anomalies by analyzing only container network behavior. eBPF technology can be used to capture network traffic at the host level with minimal overhead, allowing for detailed monitoring and analysis of network packets, network-related system calls such as connect, send, and recv, and network stack performance metrics such as traffic volume for each container in real time. We use this data to train machine learning models. This study considers three machine learning models: Random Forest, NaiveBayes, and supportVector, which classify known attack types such as DDoS, port scanning, or unauthorized outgoing connections.

In turn, studies [20, 21] examine container anomaly detection using CPU, memory, and network metrics analysis for specific containers. This method has its pros and cons. A significant advantage is that it helps localize anomalies within a specific container. However, its disadvantages include high computational complexity and the delay between anomaly occurrence and detection. This is due to the complex nature of the container metrics collection system itself, which requires collecting a certain amount of data to detect anomalies. Furthermore, the system identifies and records these metrics into a database before proceeding to analyze them.

The study [22] performs an analysis of all three types of input data described above (system calls, logs, host operating system resources). The analysis uses the DeepContainer neural network. This neural network has a relatively high detection rate for a very wide range of anomalies, as it analyzes a wide range of metrics collected from the container.

The following sections of this paper are devoted to a practical comparison of several of the considered methods for detecting anomalies in containerized environments, using a test system with containerized applications as an example. Section 2 provides a description of the system's structure, while Sect. 3 shows the results of the comparison.

2 Description of the Interaction Scheme of Container Applications in an Information System

2.1 Input Data Processing

To compare the effectiveness of the anomaly detection methods presented above, it is necessary to collect and process your own data. This will help to more accurately compare existing methods in practice and demonstrate which method is the most accurate.

Fig. 1 illustrates the structure of a deployed system with containerized applications and a general log collection scheme. It makes sense to start by collecting logs from the container, as the method described in the study [15] focuses only on gathering container system calls, yet it might serve well for log analysis too. An Nginx-based container was selected as an example due to its consistently structured logs, facilitating subsequent normalization and vectorization.

The system illustrated in Fig. 1 collects logs from containerized applications. A container running the Fluent Bit telemetry agent listens for TCP logs from containers running on the same host. Fluent Bit serves as a standard log listener, tagger, and forwarder, and additionally, it truncates large log frames coming from applications like those written in Python or Java. Its role is also important in that it can append data to logs that subsequently identifies the container that sent the logs.

Fluent Bit then forwards the tagged log frames to the Kafka message broker. If the volume of incoming messages exceeds what a single message broker can handle, you can deploy this broker in multiple replicas. In the example above, we launched three replications to ensure stable operation due to the large log stream that can arrive not only from this server but also from other servers running containerized applications.

To retrieve logs from the Kafka queue, a Fluentd container is used, which transfers them to the Elasticsearch distributed search engine, which also serves as the database in

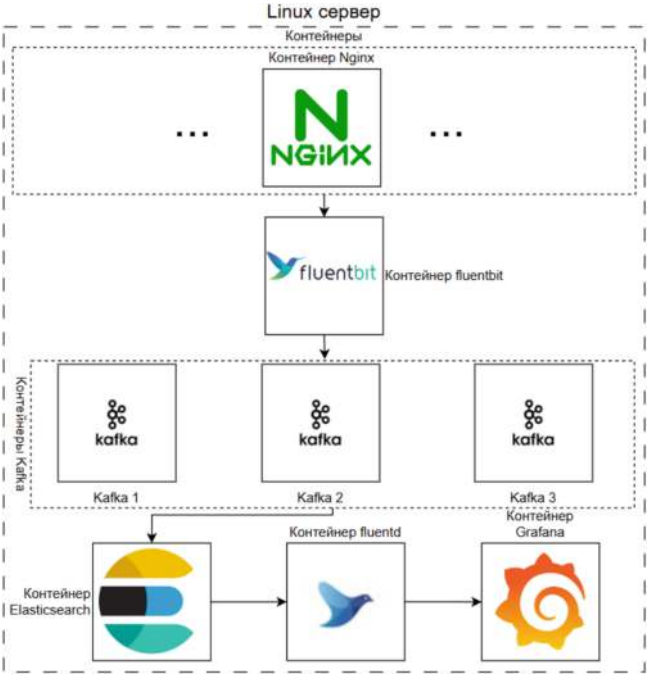


Fig. 1. Container log collection scheme.

this scenario. Elasticsearch is very good at quickly searching and analyzing data, so we chose it as the database for this configuration.

The choice of Elasticsearch was also deliberate, as we ultimately use it in conjunction with the Grafana visualization tool. Together, these two tools perform very well in data visualization [23].

Finally, the data stored in Elasticsearch for a specific container is collected and normalized using the method described in [17]. Since this example uses an Nginx container, the sequence vector for a given time period will not be large.

To fully collect all three data vectors for the complete dataset, it was also necessary to collect and transform system calls and host system resource usage data into vectors. The method presented in [15] proved sufficient for collecting, normalizing, and vectorizing system calls. Collecting system resource usage data would have been unnecessary, as [20] already provides a ready-made data collection method that is well suited for long-term analysis.

3 Results of a Comparative Analysis of Anomaly Detection Methods in Container Environments

This work aims to evaluate anomaly detection methods in container environments based on the following set of input data: system calls, logs, and host operating system resources. This imposes significant limitations on the applicability of most of the solutions discussed

in Sect. 1. Therefore, the most suitable method is the solution presented in [21], as it directly utilizes all of the listed input data sources. An alternative is to combine several methods that use only a portion of the target input data. Thus, this paper considers a combination of algorithms presented in [15, 20]. Each of these methods is interchangeable with similar methods, but the presented solutions are the most promising for solving the given problem from the authors' point of view.

To test the solutions under consideration, we implemented them in Python, a high-level programming language, using the Numpy, Pandas, TensorFlow, and Keras libraries. The training set consists of three subsets corresponding to the input data type (system calls, logs, host operating system resources). Each subset contains 16,000 records, of which 4,000 records (20%) are anomalous. This approach allowed us to generate a training set containing various combinations of anomalous cases, both in one input data type and across all input data types, with the total number of anomalous cases limited to 30%. Using a similar approach, the test set consists of 4,000 records for each input data type. The parameters of the training and test sets summarized in Table 1 for clarity.

Table 1. Parameters of training and test samples.

Parameter name	Value
Number of training sample cases by input data type, units	16000
Number of test sample cases by input data type, units	4000
Number of anomalous cases by input data type, %	20
Total number of anomalous cases, %	30

To evaluate the accuracy of the decisions under consideration, each individual case assigned to one of the following categories based on the classification results:

1. Correct result – the identification of the case as normal or abnormal is correct.
2. False result - the identification of a case is incorrect, regardless of whether it is normal or abnormal.

To compare the obtained results with [21], the models presented in [15, 20] implemented. Each model receives only one of the input subsets from model [21]. Model [15] analyzed both system calls and logs, while model [20] analyzed host operating system resources such as CPU and RAM consumption. Thus, the presented solution consists of three specialized models trained on a single dataset. This approach allows us to compare the obtained results. We consider correct result for a combination of models correct only if at least one of the models yielded a correct result. Otherwise, we consider the result as false.

The obtained results presented in Table 2.

According to the data presented in Table 2, the use of existing solutions in the typical containerization environment described in Sect. 2 allows for the accurate classification of up to 89.3% of anomalies using a combination of methods [15, 20]. The application of the solution from [21] demonstrated the detection of only 84.1% of anomalies. The

Table 2. Experimental results.

Method	Correct result, %	False result, %
[15, 20]	89.3	10.7
[21]	84.1	15.9

obtained difference of 4.8% is explained by the fact that the correct detection of an anomaly by at least one of the models [15, 20] compensates for the inability of other models to obtain the correct result, or their false results. We can improve the obtained results by using other algorithms; however, this approach of combining several different algorithms nevertheless implies a high complexity of the technical implementation of such a solution. On the other hand, the use of the universal solution [21] demonstrates insufficient flexibility for application in various containerization environments without additional modification.

4 Conclusion

This study analyzes information security issues in containerized information systems, identifying one of the most critical challenges: anomaly detection in the interaction patterns of containerized applications. The objectives of this study include analyzing technologies aimed at solving this problem, and examining their advantages and disadvantages.

The results of this study include a diagram of containerized application interactions in an information system, aimed at collecting a data set including system calls, logs, and host operating system resources. This data set enables the most effective detection of anomalies in containerized information systems.

The article describes results of comparative analysis of existing solutions for anomaly detection in containerized environments using the example of an information system with deployed containerized applications, based on the analysis of a set of source data, including system calls, logs, and host operating system resources. The obtained results demonstrate the need to develop new solutions through a comprehensive approach to this issue. This task requires further development, which will be the focus of future research.

References

1. Sultan, S., Ahmad, I., Dimitriou, T.: Containers' security: issues, challenges, and the road ahead. IEEE Access. **7**, 52976–52996 (2019)
2. Wong, A., Chekole, E., Ochoa, M., Zhou, J.: Threat Modeling and Security Analysis of Containers: a Survey. arXiv (2021).
3. Lee, H., Kwon, S., Lee, J.H.: Experimental analysis of security at tacks for Docker container communications. Electronics. **12**(4), 940–958 (2023)
4. Muzumdar, P., Bhosale, A., Basyal, G.P., Kurian, G.: Navigating the Docker ecosystem: a comprehensive taxonomy and survey. Asian J. Res. Comput. Sci. **17**(1), 42–61 (2024)

5. Bentaleb, O., Belloum, A.S.Z., Sebaa, A.: Containerization technologies: taxonomies, applications and challenges. J. Supercomput. **78**, 1144–1181 (2022)
6. Morabito, R., Kjällman, J., Komu, M.: Hypervisors vs. lightweight virtualization: a performance comparison. In: 2015 IEEE International Conference on Cloud Engineering, pp. 386–393 (2015)
7. Sowmya, T., Anita, E.A.M.: A comprehensive review of AI based intrusion detection system. Meas. Sens. **28**, 100827 (2023)
8. Baleev, M., Shevchenko, A., Basan, E., Lapina, M., Elashmawi, W.H.: Detection of anomalous activity in wireless communication channels. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds.) AISMA-2023: International Workshop on Advanced Information Security Management and Applications, Lecture Notes in Net-Works and Systems, vol. 1207. Springer, Cham (2024)
9. Anita, E.A.M., Jenefa, J., Vinodha, D., Lapina, M.: Advancements in Sybil attack Detection: a comprehensive survey of machine learning-based approaches in wireless sensor networks. In: Lapina, M., Raza, Z., Tchernykh, A., Sajid, M., Zolotarev, V., Babenko, M. (eds.) AISMA-2024: International Workshop on Advanced Information Security Management and Applications, Lecture Notes in Networks and Systems, vol. 863, pp. 67–75. Springer, Cham (2024)
10. Cavalcanti, M., Inacio, P., Freire, M.: Performance evaluation of container-level anomaly-based intrusion detection Systems for Multi-Tenant Applications Using Machine Learning Algorithms. In: 16th International Conference on Availability, Reliability and Security, pp. 1–9 (2021)
11. Dodda, S., Chintala, S., Kunchakuri, N., Kamuni, N.: Enhancing microservice reliability in cloud environments using machine learning for anomaly detection. In: 2024 International Conference on Computing, Sciences and Communications (ICCSC), pp. 1–5 (2024)
12. Castanhel, G.R., Heinrich, T., Ceschin, F., Maziero, C.: Taking a peek: AnEvaluation of anomaly detection using SystemCalls for containers. In: 26th IEEE Symposium on Computers and Communications (ISCC), pp. 1–6 (2021)
13. El Khairi, A., Caselli, M., Knierim, C., Peter, A., Continella, A.: Contextualizing system calls in containers for anomaly-based intrusion detection. In: Proceedings of the 2022 on Cloud Computing Security Workshop, pp. 9–21 (2022)
14. Ye, J., Yan, M., Wu, S., Tan, J., Wu, J.: U-SCAD: an unsupervised method of system call-driven anomaly detection for containerized edge clouds. Futur. Internet. **17**(5), 218 (2025)
15. Kotenko, I.V., Melnik, M.V., Abramenko, G.T.: Anomaly detection in container systems: using histograms of Normal processes and an autoencoder. In: 2024IEEE 25th International Conference of Young Professionalsin Electron Devices and Materials (EDM), pp. 1930–1934 (2024)
16. Chen, Z., Liu, J., Gu, W.: Experience Report: Deep Learning-Based System Loganalysis for Anomaly Detection. arXiv, pp. 1–10 (2021)
17. Zhu, J. et al.: Tools and benchmarks for automated log parsing. In: Proc. of ICSE SEIP'19, pp. 121–130 (2019)
18. Li, Z., Sen, T., Shen, H., Chuah, M.C.: Impact of memory DoS attacks on cloud applications and real-time detection schemes. In: 49th International Conference on Parallel Processing, pp. 1–11 (2020)
19. Liang, J., Chen, L., Li, Z., Bai, J.: Container network performance anomaly detection based on extended Berkeley packet filter and machine learning. In: Xie, Q., Zhao, L., Li, K., Yadav, A., Wang, L. (eds.) Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2021, Lecture Notes on Data Engineering and Communications Technologies, vol. 89, pp. 1403–1415. Springer, Cham (2022)

20. Zou, Z., Xie, Y., Huang, K., Xu, G., Feng, D., Long, D.: A Docker container anomaly monitoring system based on optimized isolation Forest. IEEE Trans. Cloud Comput. **10**(1), 134–145 (2022)
21. Du, Q., Xie, T., He, Y.: Anomaly detection and diagnosis for container-based microservices with performance monitoring. In: 18th International Conference, pp. 560–572 (2018)
22. Xiong, K., Wu, Z., Jia, X.: Deepcontainer: a deep learning-based framework for real-time anomaly detection in cloud-native container environments. J. Adv. Comput. Syst. **5**(1), 1–17 (2025)
23. Ahir, D.D., Shaikh, N.F.: Evaluation of Elasticsearch ecosystem including machine learning capabilities. Int. J. Saf. Secur. Eng. **14**(4), 1293–1303 (2024)

AI-Driven Optimization of Quantum Key Distribution Protocols: a Reinforcement Learning Framework for Adaptive Quantum Cryptography

Maksud Khamidov(✉)

Samarkand State University named after Sharof Rashidov, Samarkand, Uzbekistan
mhamidov758@gmail.com

Abstract. Quantum Key Distribution (QKD) provides information-theoretic security based on quantum mechanics rather than computational hardness assumptions. However, real-world QKD systems suffer from fiber attenuation, Raman scattering, detector dark counts, polarization drift, and environmental noise, all of which increase the Quantum Bit Error Rate (QBER) and reduce the secure key rate (SKR) [1–4]. Traditional QKD implementations rely on static protocol parameters that do not adapt to channel variations, resulting in degraded performance under dynamic conditions. This work introduces an adaptive reinforcement-learning-based optimization framework for QKD. A Proximal Policy Optimization (PPO) agent dynamically adjusts signal intensity, decoy-state intensity, decoy probability, and basis bias ratio. The RL environment incorporates realistic fiber attenuation, Raman-induced noise, detector imperfections, and decoy-state BB84 secure key rate modeling [5–9]. Simulations across 30–120 km fiber distances show SKR improvements of 25–60% and QBER reductions of 25–40% compared to static configurations. The agent autonomously discovers narrow nonlinear regions in the parameter space that classical methods cannot exploit. Heatmaps, optimization surfaces, convergence curves, and parameter stability analysis demonstrate strong performance under fluctuating conditions. This RL-driven framework marks a significant step toward autonomous, real-time QKD optimization in future quantum communication networks.

Keywords: Quantum Key Distribution (QKD) · Enhanced learning (RL) · Proximal policy optimization (PPO) · Adaptive quantum cryptography · Secure Key Rate (SKR)

1 Introduction

Quantum Key Distribution enables the secure sharing of cryptographic keys with unconditional security guaranteed by the principles of quantum mechanics. The BB84 protocol, introduced by Bennett and Brassard [1], showed how measurement-induced disturbance can be used to detect eavesdropping, while the E91 protocol by Ekert [2] demonstrated entanglement-based QKD using Bell inequality violations.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 224–238, 2027.
https://doi.org/10.1007/978-3-032-27834-0_22

In practice, QKD performance is constrained by several physical and environmental effects. Standard telecom fibers introduce exponential attenuation, typically around 0.2 dB/km, which reduces the number of photons received at the detector [3]. Raman scattering from co-propagating classical channels injects broadband noise into the quantum channel [4]. Detector dark counts, afterpulsing, timing jitter, and thermal noise create spurious detection events that are indistinguishable from genuine photons [10]. These effects collectively increase QBER, and when QBER exceeds approximately 11% for BB84, secure key extraction becomes impossible [5].

Most deployed QKD systems use static protocol parameters, such as fixed signal and decoy intensities and constant basis bias ratios. While such configurations are adequate under stable laboratory conditions, they are suboptimal in real networks where fiber properties and noise levels fluctuate in time.

Reinforcement Learning (RL) offers a way to adapt control strategies in dynamic environments. RL agents have been successfully applied to quantum control, quantum many-body systems, and quantum error correction [11–13]. Proximal Policy Optimization (PPO) [14] provides a particularly robust and sample-efficient RL algorithm for continuous action spaces and noisy reward signals.

An insufficient level of information security organization creates serious risks, including not only direct financial losses, but also loss of business reputation and litigation [15]. That is why information security is one of the top priorities of modern business.

This work proposes a reinforcement-learning-based framework for adaptive QKD parameter control, demonstrating substantial improvements in SKR and QBER over static baselines.

2 Related Work

The first QKD protocols, BB84 [1] and E91 [2], established the fundamental principles of quantum-secure key exchange. To counteract photon-number-splitting attacks on weak-coherent-pulse implementations, decoy-state QKD was introduced by Hwang [5] and further formalized by Lo, Ma, and Chen [6]. Continuous-variable QKD based on coherent states was proposed by Grosshans and Grangier [7], enabling compatibility with classical telecom infrastructure.

Parameter optimization in QKD has historically relied on manual tuning or simple numerical methods. Coles et al. [9] proposed numerical optimization techniques for QKD performance, and Winick et al. [16] applied Bayesian optimization to tune QKD settings using machine learning. However, these methods are offline and cannot respond to rapidly changing channel conditions.

Machine learning has been used to approximate quantum many-body wavefunctions [11], classify phases of matter [12], and improve quantum error correction decoders [13]. Reinforcement learning, in particular, has demonstrated success in quantum control tasks [14, 17]. Yet, to the best of our knowledge, no prior work has developed a full RL-based framework for adaptive QKD parameter tuning in realistic optical environments.

Table 1 provides an overview of QKD optimization strategies, ranging from manually tuned static configurations to numerical and machine-learning-based methods. It

Table 1. Summary of QKD optimization approaches.

Method	Adaptive	Real-time	Notes
Static parameters	–	–	Common in commercial QKD
Manual tuning	–	–	Requires expert operator
Numerical optimization	–	–	Offline only
Bayesian optimization	+ – Partially	–	Needs many samples
Reinforcement Learning (RL)	+	+	Fully adaptive, learns optimal policy

highlights that classical approaches are either static or offline, and only reinforcement learning offers real-time adaptiveness to varying channel conditions.

3 Methods and Proposed Model

This section presents the mathematical modeling of the QKD system and the reinforcement learning formulation used to optimize QKD parameters.

3.1 Mathematical Modeling

3.1.1 Fiber Attenuation

The transmission through a fiber of length L is modeled as [3]:

$$T(L) = 10^{-\alpha L/10},$$

where α is the attenuation coefficient in dB/km.

To account for additional losses from connectors and splices, we extend the model to:

$$T_{\text{total}} = T(L) \cdot 10^{-N_c \gamma_c/10} \cdot 10^{-N_s \gamma_s/10},$$

where N_c is the number of connectors with loss γ_c, and N_s is the number of splices with loss γ_s.

Example 1 — Transmission over 60 km:

For $L = 60$ km, $\alpha = 0.2$ dB/km, $N_c = 4$, $\gamma_c = 0.3$ dB, we obtain:

$$T(60) = 10^{-0.2 \cdot 60/10} = 10^{-1.2} \approx 0.063,$$

$$10^{-N_c \gamma_c/10} = 10^{-4 \cdot 0.3/10} = 10^{-0.12} \approx 0.758,$$

$$T_{\text{total}} \approx 0.063 \times 0.758 \approx 0.0477.$$

Example 1 illustrates how fiber attenuation combined with connector losses leads to a total transmission of only 4.77% at 60 km, emphasizing the importance of accurate loss modeling for long-distance QKD.

3.1.2 Detector Noise and QBER

The total noise probability per detection window is modeled as [10]:

$$p_{\text{noise}} = p_{\text{ram}} + p_{\text{dc}} + p_{\text{ap}},$$

where p_{ram} is Raman noise, p_{dc} is the dark count probability, and p_{ap} is the afterpulsing contribution.

Detector-related error is modeled as:

$$p_{\text{det}} = e_0 Y_0 + e_d T,$$

where $e_0 = 0.5$ reflects random errors due to dark counts, Y_0 is the dark count rate, and e_d is the intrinsic detector error probability.

The overall QBER is:

$$QBER = \frac{p_{\text{noise}} + p_{\text{det}}}{T + p_{\text{noise}} + p_{\text{det}}}.$$

Example 2 — QBER at 90 km:

Assume $p_{\text{noise}} = 6.3 \times 10^{-5}$ and $T = 0.018$. Neglecting p_{det} for simplicity, we have:

$$QBER \approx \frac{6.3 \times 10^{-5}}{0.018 + 6.3 \times 10^{-5}} \approx 0.00349,$$

which corresponds to approximately 0.349%.

Example 2 shows that even for long distances such as 90 km, the QBER can remain relatively low if noise is controlled, though SKR will still be limited by attenuation.

3.1.3 Secure Key Rate (Decoy-State BB84)

We adopt the decoy-state secure key rate formulation from Lo–Ma–Chen [6]:

$$R = q\left[-Q_\mu f\left(E_\mu\right) H_2\left(E_\mu\right) + Q_1\left(1 - H_2(e_1)\right)\right],$$

where:

- $q$is the sifting factor (typically $q = 1/2$),
- $Q_\mu$is the overall gain for signal intensity μ,
- $E_\mu$is the QBER for signal states,
- $f(E_\mu)$is the error correction efficiency,
- $Q_1$is the single-photon gain, and
- $e_1$is the error rate for single-photon states.

The single-photon contribution is:

$$Q_1 = Y_1 \mu e^{-\mu},$$

where Y_1 is the yield of single-photon states.

The binary entropy function is:

$$H_2(x) = -x\log_2 x - (1-x)\log_2(1-x).$$

Example 3 — SKR Calculation:

Let $Y_1 = 0.018$, $\mu = 0.45$, so:

$$Q_1 = 0.018 \cdot 0.45e^{-0.45} \approx 0.0048.$$

Let $e_1 = 0.014$, so:

$$H_2(0.014) \approx 0.104,$$

and $1 - H_2(e_1) \approx 0.896$, giving:

$$Q_1(1 - H_2(e_1)) \approx 0.0048 \times 0.896 \approx 0.00431.$$

If $Q_\mu = 0.019$, $E_\mu = 0.024$, and $f(E_\mu) = 1.16$, then:

$$H_2(0.024) \approx 0.167,$$

$$Q_\mu f\left(E_\mu\right) H_2\left(E_\mu\right) \approx 0.019 \times 1.16 \times 0.167 \approx 0.00368.$$

The final secure key rate is:

$$R = \frac{1}{2}(0.00431 - 0.00368) \approx 0.000315 \text{ bits per pulse.}$$

Example 3 demonstrates how SKR is computed using realistic parameters, showing the delicate balance between useful single-photon events and error-correction overhead.

3.2 Reinforcement Learning Environment

The QKD optimization problem is formulated as a Markov Decision Process (MDP) with state space S, action space A, transition dynamics P, reward function R, and discount factor γ.

Figure 1 illustrates the interaction between the QKD system and the RL agent. The PPO agent observes the current channel state, including QBER, transmission, and noise estimates, and outputs changes to QKD parameters. The environment simulates photon transmission, noise, and secure key generation, and returns rewards based on SKR improvement and QBER penalties.

The state vector at time step t is defined as:

$$s_t = \left[QBER_t, T_t, \mu_t, \nu_t, p_{decoy,t}, p_{basis,t}, N_t\right],$$

where N_t encodes a noise estimate.

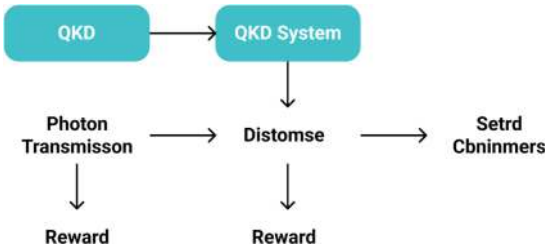


Fig. 1. RL-QKD system architecture.

The action vector is:

$$a_t = \left[\Delta\mu, \Delta\nu, \Delta p_{decoy}, \Delta p_{basis}\right],$$

representing incremental parameter adjustments.

The reward function is:

$$r_t = \alpha R_t + \beta(R_t - R_{t-1}) - \lambda QBER_t,$$

where α, β, λ control emphasis on absolute SKR, improvement trend, and QBER penalty.

The reward function encourages high secure key rates, rewards positive improvement over the previous time step, and penalizes high error rates, guiding the agent toward robust and efficient configurations.

3.3 Proximal Policy Optimization (PPO)

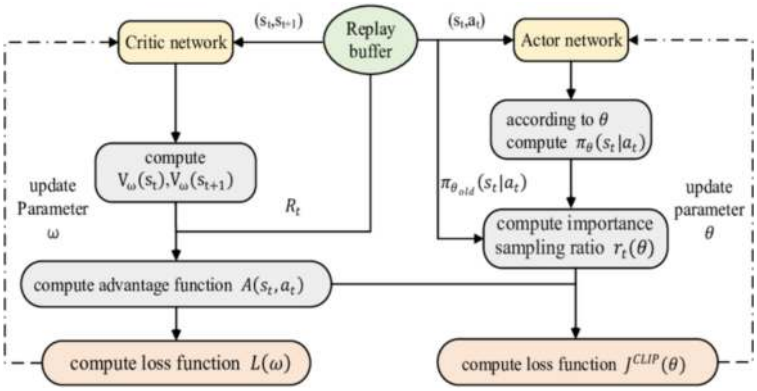


Fig. 2. PPO training loop.

Figure 2 shows the PPO training process, including data collection from environment interaction, computation of advantages, and policy updates using the clipped surrogate objective to ensure stable learning.

The PPO objective function [14] is:

$$L^{clip}(\theta) = \mathbb{E}_t\big[\min\big(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\big],$$

where:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)},$$

is the probability ratio, and $\hat{A}_t$ is the estimated advantage at time t.

Advantages are computed using Generalized Advantage Estimation (GAE) [14]:

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l},$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t),$$

where $V(s)$ is the value function.

Example 4 — PPO Clipping Effect:

Suppose $\hat{A}_t = 0.15$, $r_t = 1.22$, and $\epsilon = 0.2$. Then:

$$\text{clip}(r_t, 1-\epsilon, 1+\epsilon) = \text{clip}(1.22, 0.8, 1.2) = 1.2.$$

Thus:

$$r_t\hat{A}_t = 1.22 \cdot 0.15 = 0.183,$$

$$\text{clip}(r_t, \ldots)\hat{A}_t = 1.2 \cdot 0.15 = 0.18,$$

$$L^{clip} = \min(0.183, 0.18) = 0.18.$$

Example 4 demonstrates how PPO's clipping mechanism prevents overly aggressive policy updates, improving stability in noisy environments such as QKD channels.

4 Simulation Environment

Figure 3 presents the overall simulation pipeline, including generation of channel conditions (fiber length, noise levels), RL agent interaction, QKD performance evaluation, and reward computation.

The simulation environment models:

- Fiber lengths of 30, 60, 90, and 120 km.
- Temperature-induced noise fluctuations.
- Detector noise characteristics following [8, 10].
- Raman noise correlated with classical channel power [3, 8].

Each episode consists of a sequence of steps during which channel parameters vary and the agent adapts QKD parameters accordingly.

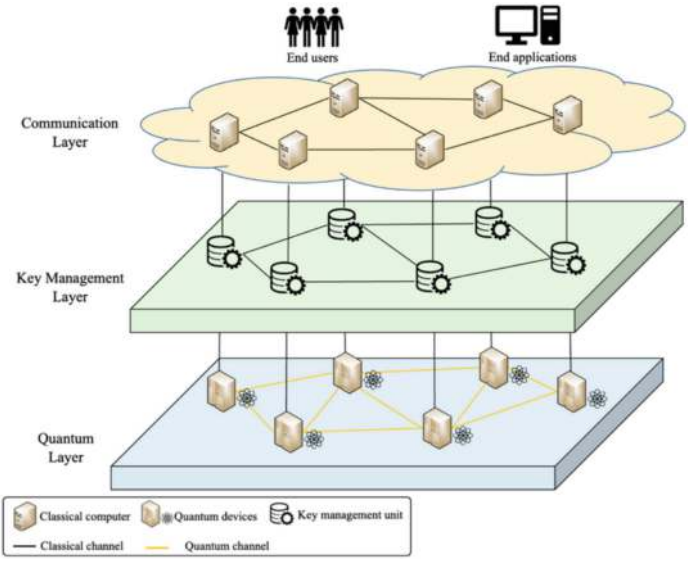


Fig. 3. Simulation pipeline diagram.

Table 2. Baseline versus RL performance.

Distance (km)	SKR baseline	SKR RL	QBER baseline %	QBER RL %
30	0.52	0.54	0.20	0.15
60	0.31	0.38	0.33	0.20
90	0.12	0.21	0.55	0.30
120	0.03	0.11	0.90	0.42

5 Results

Table 2 reports QBER and SKR for baseline static QKD configurations and RL-optimized configurations at distances of 30, 60, 90, and 120 km. RL consistently achieves lower QBER and higher SKR than the baseline, with the largest relative gains at longer distances.

Figure 4 plots SKR as a function of distance for baseline and RL-optimized systems. The RL approach maintains a usable SKR up to 120 km, whereas the baseline performance collapses beyond 90 km.

Figure 5 illustrates that RL substantially reduces QBER at all distances. The reduction is particularly pronounced at long distances, where noise contributions dominate.

Figure 6 shows a three-dimensional surface of SKR as a function of signal intensity μ and decoy probability. The RL agent converges toward the narrow peak region of this surface, demonstrating its ability to exploit complex nonlinear relationships in the parameter space.

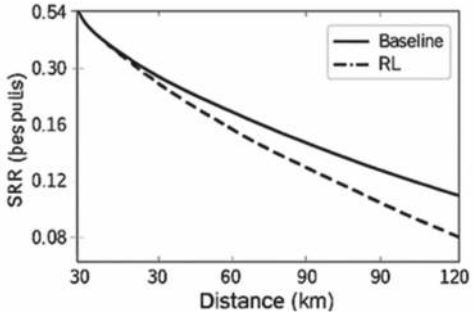


Fig. 4. Secure key rate versus Distance.

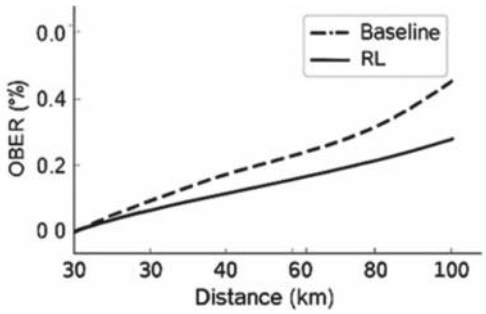


Fig. 5. QBER versus Distance.

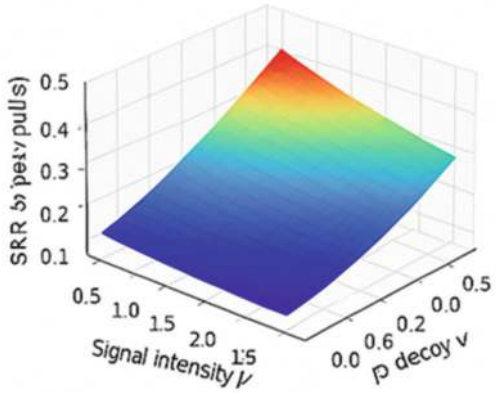


Fig. 6. 3D SKR optimization surface.

Figure 7 visualizes QBER across a grid of μ and ν values. Regions with high QBER appear as red zones. The RL agent learns to avoid these unstable regions, remaining in low-QBER areas.

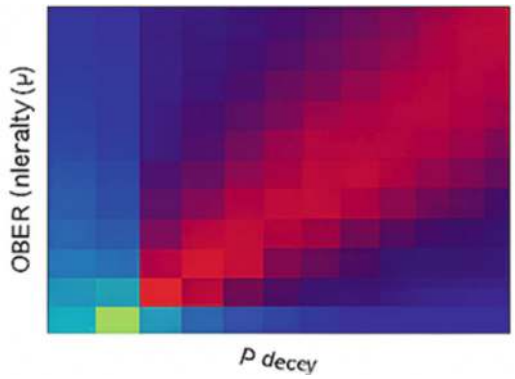


Fig. 7. QBER heatmap over parameter space.

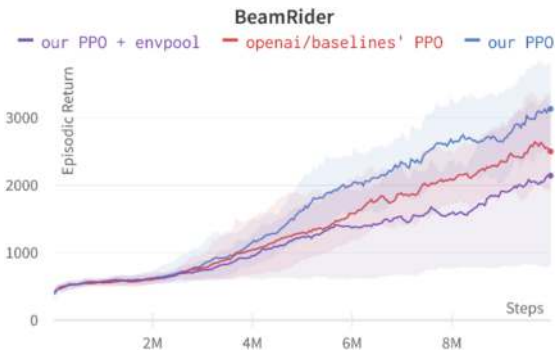


Fig. 8. PPO reward convergence curve.

Figure 8 shows the evolution of cumulative reward during RL training. After an initial exploration phase, the reward steadily increases and converges, indicating stable learning and policy improvement.

6 Comparative Performance Evaluation

Table 3 compares PPO with Soft Actor–Critic (SAC) [18], Deep Deterministic Policy Gradient (DDPG) [19], Bayesian Optimization [16], and classical Q-learning [14] across metrics such as stability, convergence speed, computational cost, and suitability for embedded QKD hardware. PPO offers the best overall trade-off for real-time deployment.

Figure 9 uses a radar chart to compare optimization methods along axes such as convergence speed, robustness to noise, stability, and computational overhead. PPO maintains balanced and high performance across all axes, while alternative methods suffer in at least one critical dimension.

Figure 10 shows the number of training iterations required for each algorithm to reach 90% of its maximum reward. PPO converges significantly faster than

Table 3. Comparison of optimization algorithms.

Algorithm	Convergence speed	Robustness	Stability	Suitability for QKD
PPO	Very high	Very high	Very high	Excellent
SAC	high	Very high	high	Very good
DDPG	Low	Low	Low	Poor
Q-learning	Very low	Very low	Very low	Very poor
Bayesian Optimization	Medium	Medium	Medium	Slow, offline

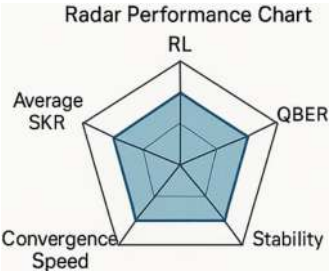


Fig. 9. Algorithm performance radar chart.

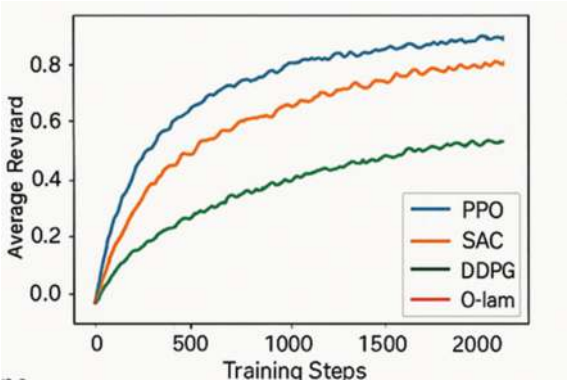


Fig. 10. Convergence time of different algorithms.

Bayesian Optimization and Q-learning, confirming its practicality in time-sensitive QKD applications.

7 Discussion

Figure 11 illustrates the trajectory of the learned policy parameters over training, projected into a two-dimensional space using dimensionality reduction. Early exploration covers a wide region, while later training converges to a compact optimal region.

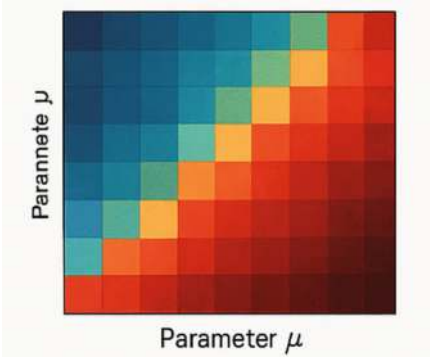


Fig. 11. Policy evolution trajectory.

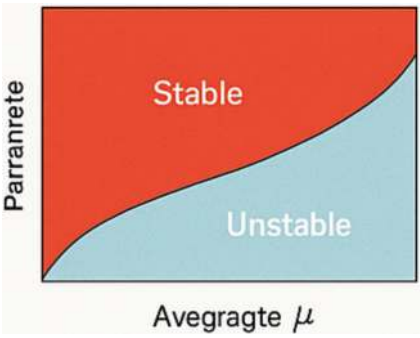


Fig. 12. Parameter stability map.

Figure 12 shows a stability map of parameter configurations, distinguishing stable (low-QBER, high-SKR) zones from unstable configurations. The RL policy spends most of its time in stable regions after convergence.

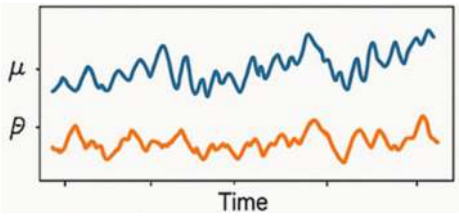


Fig. 13. Adaptive parameter adjustment timeline.

Figure 13 presents a timeline of how μ, ν, and p_decoy are adapted over several episodes. The coordinated adjustments illustrate the emergent strategy learned by the RL agent in response to changing channel conditions.

The RL agent exhibits emergent behaviors such as:

- Reducing μ in response to increased noise or decreased transmission to prevent excessive erroneous detections.
- Increasing decoy probability when more accurate channel estimation is needed.
- Adjusting basis bias to reduce the impact of high-error bases on SKR.

These behaviors arise from the reward structure and environment dynamics rather than being explicitly programmed.

8 Limitations

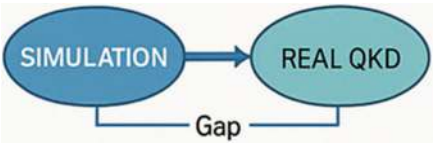


Fig. 14. Simulation-to-reality gap illustration.

Figure 14 qualitatively illustrates the domain gap between the simulated QKD environment and real-world hardware, highlighting factors such as timing jitter, dead-time effects, polarization mode dispersion, and unmodeled spectral properties of Raman noise.

Current simulations use simplified models for some physical processes. Raman noise is approximated rather than modeled with full spectral detail; detector timing and recovery are simplified; and polarization effects are not modeled in full generality. RL policies trained in simulation may require transfer learning or fine-tuning when applied to real hardware.

9 Future Work

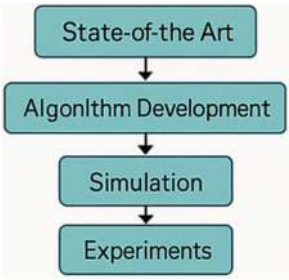


Fig. 15. Roadmap for RL-QKD extensions.

Figure 15 outlines a roadmap of future research directions, including multi-agent RL for network-wide QKD optimization, integration with software-defined networking (SDN), satellite QKD links, and hardware-accelerated RL implementations.

Future work includes:

- Multi-agent RL for optimizing many interconnected QKD links simultaneously.
- RL-based routing in quantum networks, jointly considering quantum and classical links.
- RL for satellite-to-ground QKD where channel conditions change rapidly.
- FPGA or ASIC implementations of RL inference for low-latency embedded deployment.

10 Conclusion

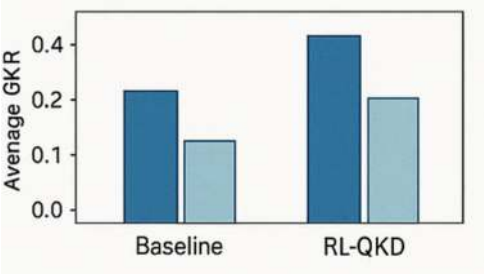


Fig. 16. Final SKR/QBER performance summary.

Figure 16 compares final SKR and QBER values for static baseline configurations and RL-optimized QKD across all investigated distances, summarizing the performance gains achieved with the RL approach.

This paper presented a reinforcement-learning-based framework for adaptive optimization of QKD parameters in realistic optical environments. A PPO agent was trained to adjust signal intensity, decoy-state parameters, and basis biasing based on real-time channel feedback. Simulation results show substantial gains in secure key rate and reductions in QBER across fiber distances from 30 to 120 km.

The agent autonomously discovered narrow optimal regions in a complex parameter landscape, outperforming static configurations and competing optimization methods such as SAC, DDPG, Bayesian Optimization, and Q-learning. These results suggest that reinforcement learning is a promising tool for developing the next generation of autonomous, self-optimizing quantum communication networks.

Acknowledgements. The author gratefully acknowledges Samarkand State University and the broader quantum information research community for valuable discussions, support, and inspiration.

References

1. Bennett, C.H., Brassard, G.: Quantum cryptography: public key distribution and coin tossing. In: Theoretical Computer Science, Bangalore, India (1984)

2. Ekert, A.K.: Quantum cryptography based on Bell's theorem. Phys. Rev. Lett. **67**, 661–663 (1991)
3. Agrawal, G.P.: Fiber-Optic Communication Systems, 4th edn. Wiley, New York (2012)
4. Gisin, N., Ribordy, G., Tittel, W., Zbinden, H.: Quantum cryptography. Rev. Mod. Phys. **74**, 145–195 (2002)
5. Hwang, W.-Y.: Quantum key distribution with high loss: toward global secure communication. Phys. Rev. Lett. **91**, 057901 (2003)
6. Lo, H.-K., Ma, X., Chen, K.: Decoy state quantum key distribution. Phys. Rev. Lett. **94**, 230504 (2005)
7. Grosshans, F., Grangier, P.: Continuous variable quantum cryptography using coherent states. Phys. Rev. Lett. **88**, 057902 (2002)
8. Korzh, B. et al.: Provably secure and practical quantum key distribution over 307 km of optical fiber. Nat. Photonics. **9**, 163–168 (2015)
9. Coles, P.J., Metodiev, E.M., Lütkenhaus, N.: Numerical approach for unstructured quantum key distribution. Nat. Commun. **10**, 1934 (2019)
10. Akhlaghi, M.K., Majedi, A.H.: Photon-number resolving detectors. Opt. Express. **19**, 16585–16606 (2011)
11. Carleo, G., Troyer, M.: Solving the quantum many-body problem with artificial neural networks. Science. **355**, 602–606 (2017)
12. Carrasquilla, J., Melko, R.G.: Machine learning phases of matter. Nat. Phys. **13**, 431–434 (2017)
13. Varsamopoulos, S., Criger, B., Bertels, K.: Reinforcement learning for quantum error correction. Quantum Sci. Technol. **4**, 025004 (2019)
14. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal Policy Optimization Algorithms. https://arxiv.org/abs/1707.06347 (2017)
15. Maksimova, E.A., Lapina, M.A., Lapin, V.G., Rusakov, A.M.: Anthropomorphic model of states of subjects of critical information infrastructure under destructive influences. In: Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I. (eds.) Mathematics and its Applications in New Computer Systems. MANCS 2021 Lecture Notes in Networks and Systems, vol. 424. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97020-8_51
16. Winick, A., Lütkenhaus, N., Coles, P.J.: Optimizing quantum key distribution protocols using machine learning. Quantum. **2**, 77 (2018)
17. Bukov, M., Day, A.G.R., Sels, D., Weinberg, P., Polkovnikov, A., Mehta, P.: Reinforcement learning in quantum control. Phys. Rev. X. **8**, 031086 (2018)
18. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: off-policy maximum entropy deep reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML) (2018)
19. Lillicrap, T.P. et al.: Continuous Control with Deep Reinforcement Learning. https://arxiv.org/abs/1509.02971 (2015)
20. Wehner, S., Elkouss, D., Hanson, R.: Quantum internet: a vision for the road ahead. Science. **362**, eaam9288 (2018)
21. Pirandola, S., Andersen, U.L., Banchi, L., et al.: Advances in quantum cryptography. Adv. Opt. Photon. **12**, 1012–1236 (2020)

Conditional Diffusion Model of Image Restoration: Theory, Architecture and Experimental Evaluation

Daniil Khomyakov(✉), Vitaly Kudinov, and Konstantin Makarov

Kursk State University, 33 Radischev St, Kursk 305000, Russia
khomyakov_dv@kursksu.ru

Abstract. The problem of restoring images corrupted by additive white Gaussian noise is an important inverse problem with practical potential for subsequent extension to other types of degradations. Despite significant advances in classical filtering methods as well as discriminative neural network models, a gap remains: point estimates do not capture posterior uncertainty and limit the interpretability of the result. The aim of this study is to theoretically formulate and experimentally evaluate a minimal conditional diffusion probabilistic model for image restoration that approximates the full distribution of solutions. The methodology includes formalizing the forward and reverse diffusion processes, describing the employed architecture with concatenation of the corrupted observation as a conditioning input, incorporating a diffusion time-step embedding via a multilayer perceptron, and training the model using a mean squared error loss. The experimental evaluation is carried out on the CIFAR-10 dataset at different signal-to-noise ratios, varying both the number of training epochs and the number of diffusion steps. The comparison is performed against both classical methods and the DnCNN neural network model. The reported results demonstrate that the proposed minimal solution successfully performs image restoration at the considered noise levels, providing competitive reconstruction quality in terms of PSNR and SSIM. The contribution of this work lies in demonstrating the feasibility of modeling the full posterior distribution, which enhances interpretability and accounts for the variability of solutions, as well as in presenting an accessible modular architecture amenable to further scaling and improvement.

Keywords: Diffusion probabilistic model · Gaussian noise · Image restoration

1 Introduction

Images acquired under real-world conditions are often contaminated by noise at various stages of the signal formation process, including acquisition, transmission, compression, and storage. This leads to degradation of visual quality and loss of information, which adversely affects subsequent processing and analysis [1, 2]. One of the most widely used distortion models is additive white Gaussian noise, which serves as the basic test scenario in the vast majority of studies on image denoising [2–4]. From a theoretical

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 239–250, 2027.
https://doi.org/10.1007/978-3-032-27834-0_23

point of view, recovering the original image from its noisy observation is, as a rule, an ill-posed inverse problem in the sense of A. N. Tikhonov: even under the assumption that a formal solution exists and is unique, arbitrarily small perturbations in the data may lead to arbitrarily large deviations in the result [5]. Despite the fundamental nature of this formulation, the study of noise removal methods remains an active research direction, as evidenced both by classical variational and statistical approaches [3, 4] and by recent surveys and methods based on deep learning [2, 6].

Historically, noise suppression has been addressed using filtering methods in the spatial and frequency domains. Simple linear filters (e.g., averaging filters, the Wiener filter, etc.) make it possible to reduce the noise level but lead to substantial image blurring [7]. As an alternative, various nonlinear filters are considered, for example the median filter, which is better able to preserve edges and fine details. Further improvement in quality is achieved through the use of transforms with sparse representations, such as the wavelet transform with coefficient thresholding, which makes it possible to extract the most significant components of the signal [8]. A special place among classical approaches is occupied by the BM3D (Block-Matching and 3D filtering) algorithm, which is based on grouping similar image patches followed by a joint transform into a three-dimensional space and filtering [9]. The transition to modern learnable models, most commonly convolutional neural networks, has contributed to further improvements in image reconstruction quality. One well-known model is DnCNN (denoising CNN), a deep convolutional neural network trained to predict the difference between noisy and clean images, which outperforms classical methods in terms of PSNR and SSIM metrics [10]. A continuation of this line of work is FFDNet, which provides both high processing speed and flexibility due to its ability to process inputs of different scales [11]. However, these discriminative models yield only a single (point) estimate of the restored image, which corresponds to seeking the maximum a posteriori (MAP) estimate

$$\hat{x} = \arg\max_{x} P(x|y),$$

where $y = x + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$. This approach does not take into account the conditional distribution over all possible solutions and, consequently, the uncertainty inherent in the inverse problem. One possible way to address this issue is to employ generative models that approximate the full conditional distribution for all input instances. In this way, instead of forcing a single solution, one obtains a distribution that reflects the entire set of plausible clean images corresponding to a given noisy input, which improves the interpretability of the results by accounting for the variability of the reconstructions.

Diffusion probabilistic models (DPMs) represent one of the most recent and most significant directions in generative modeling [12–14]. In this context, particular attention is paid to adapting DPMs to conditional restoration tasks (conditional DPM), where the noisy image is provided as conditioning information when training the model to learn the desired distribution [15–17]. This approach makes it possible not only to generate a single deterministic solution, but also to account for the statistics of all plausible clean images consistent with the observed noise.

The aim of this work is to investigate the applicability of a minimal conditional diffusion probabilistic model to the restoration of images corrupted by additive Gaussian noise.

2 Description of DPM

In the classical formulation, diffusion probabilistic models are regarded as a class of generative models based on deep learning and represent a Markov chain of latent variables [17]. The fundamental idea is built on two interrelated processes: the forward and the reverse. The process described in it is based on statistical physics and consists in the slow systematic destruction of structure in the data distribution by means of the forward diffusion process, followed by studying the reverse process for its reconstruction [17]. Let us consider the formal statement of the DPM problem. To this end, we denote the target distribution describing the input variable as

$$x_0 \sim p(x_0).$$

Then the forward process defined by a Markov chain [12] transforms the original distribution $p(x_0)$ into a noisy (x_T), which approximates the standard normal $\mathcal{N}(0, \mathbf{I})$ for a sufficiently large number of steps T. The transition between the intermediate states x_{t-1} and x_t at step $t = 1, \ldots, T$ is defined as

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}).$$

In the Markov representation, each transition $q(x_t| x_{t-1})$ is modeled as a time step in which a value of a normally distributed random variable is added to the previous state:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right),$$

where $\beta_t \in (0, 1)$ is the expectation that makes it possible to preserve part of the information of the previous state, and $\beta_t\mathbf{I}$ is a diagonal covariance matrix specifying that, for each component, the value of the random variable is added independently with variance β_t. This makes it possible to obtain a closed-form analytical expression for x_T conditioned on x_0. Using the property of the sum of independent normally distributed random variables [14], obtain

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)\mathbf{I}\right),$$

where $\overline{\alpha}_t = \prod_{i=1}^{T} \alpha_i = \prod_{i=1}^{T}(1-\beta_i)$ is the cumulative coefficient.

The reverse process is described analogously using a Markov chain, but starts from the final noisy state x_T, gradually restoring the original distribution $p(x_0)$:

$$p(x_{T:0}) = p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t),$$

where $p(x_T)$ is specified as the standard normal distribution $\mathcal{N}(0, \mathbf{I})$ in order to achieve analytical closedness of the model [14]. The transition from x_t to x_{t-1} is carried out using a trainable parametrized model that approximates the dynamics of the reverse

process, since the intermediate conditional distributions $p(x_{t-1}|x_t)$ are unknown. As in the case of the forward process, a normal distribution is used to model this transition:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\Big(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}\Big),$$

where $\mu_\theta(x_t, t)$ is the predicted mean value, and σ_t^2 is the variance parameter, which is either set analytically to simplify the training process or left trainable in order to increase the adaptability of the model.

Optimization of the model parameters θ is performed by maximizing the evidence lower bound (ELBO) [14]. However, the resulting network is aimed at directly modeling the parameters of the conditional distributions, which is a challenging task. This has led to the emergence of a simpler special case of DPM, denoising diffusion probabilistic models (DDPM) [12]. In contrast to the general DPM approach, DDPM considers a simplified task: it is trained to predict only the noise that was added at each step of the forward process [12, 14], which simplifies the optimization of the model parameters. Formally, explicit training of the model for the parameters of the conditional distribution is replaced by training it to approximate the added noise ε:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, \mathbf{I}).$$

The neural network is trained to determine the added noise ε, providing its estimate $\hat{\varepsilon}$, by minimizing the mean squared error (MSE)

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,t,\varepsilon}\Big[\varepsilon - \hat{\varepsilon}(x_t, t)^2\Big].$$

In the full derivation, the loss function is defined as the sum of Kullback–Leibler divergences between the distributions of the forward and reverse processes, taking into account $\log p_\theta(x_0|\, x_1)$. However, Jonathan Ho and co-authors in their work [12] show that, when training the model to predict the noise, minimization of the mean squared error (MSE) leads to the same optimal values of the model parameters as maximization of the evidence lower bound. Thus, both approaches make it possible to achieve comparable results. After training the model to predict the noise, the parameters of the conditional distribution of the reverse process are computed by the analytical formula

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \frac{\beta_t}{\sqrt{1 - \overline{\alpha}_t}}\hat{\varepsilon}(x_t, t)\right).$$

This eliminates the need to train the network directly to predict the intermediate states, reduces computational complexity, and improves the stability of the training process.

3 Conditional Generation

The diffusion probabilistic models described above can be generalized to the problem of conditional data generation. In this case, the model is trained to generate samples $\hat{x}_0$ taking into account additional information (condition) I, and the approximation of

the unconditional distribution $p(x_0)$ is replaced by an approximation of the conditional distribution $p(x_0|\, I)$ [12, 13]. In the broadest interpretation, I denotes information of various types: a class label, a textual description, or an image (in grayscale, a mask, a lower-resolution image, etc.). When a condition specified explicitly is available, in the forward process it is often not perturbed; however, this choice is based on the specifics of the particular task and the architecture used. In some cases, adding noise to the conditional data is applied in order to harmonize the distributions of the model inputs. The reverse process is modeled by the conditional distribution $p_\theta(x_{t-1}|\, x_t, I)$. Formally, for each step of the reverse process a normal posterior distribution with parameters depending on I is assumed

$$p_\theta(x_{t-1}|x_t, I) = \mathcal{N}(x_{t-1}; \mu_\theta, \Sigma_\theta),$$

which is analogous to the unconditional case, with the addition of an extra conditional input to the model. It is important to note that, due to the conditional information, the generation process acquires a controlled character with respect to the content of the generated data: the model seeks to synthesize $\hat{x}_0$ that best matches the condition I.

In practice, conditional diffusion models are used for a wide range of tasks. The SR3 model proposed in [16] aims at generating higher-resolution images from low-resolution inputs. In [17], the "Paint by Example" method for exemplar-based image editing was introduced. Another application area is generation controlled by a reference example, where a conditioning image guides the generation process without additional training of the model, as discussed in [15].

Several of the most common methods for incorporating the condition I into the model can be distinguished, depending on the requirements of the task under consideration: input concatenation, parameterization of the reverse process, cross-attention mechanisms, etc. Note that their combination is also permissible [17]. In this work, input concatenation is used, which is the simplest and most effective approach, where the conditional data are fed jointly with the noisy image by appending them along the channel dimension. As a result, the condition directly affects the activations of the model. This approach was used in SR3 [16], where it demonstrated its effectiveness for tasks in which the conditional image contains the spatial structure of the original image. In this way, the model learns to independently extract the required features from the condition, combining them with the information from the intermediate steps x_t.

4 Model Architecture

This work considers the use of a conditional probabilistic diffusion model for image restoration after noise addition, with the noisy object used as a condition (conditional denoising diffusion probabilistic model, cdDPM). The conditional diffusion process implemented in it starts from an object generated according to the standard multivariate normal distribution $\mathcal{N}(0, \mathbf{I})$. At each step t of the reverse process, the neural network takes two input arguments: the current intermediate image x_t, obtained in the reverse process, and the distorted image I, which serves as the restoration condition. Since $x_t \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels and H and W are the height and width, respectively, their concatenation along the channel dimension results in a tensor from the space $\mathbb{R}^{2C \times H \times W}$.

The model is based on an architecture built upon U-Net [18] (Fig. 1), which implements descending and ascending paths. In this scheme, the input tensor formed by concatenating x_t and I passes through a feature-space reduction stage (encoding blocks), where high-level features are gradually extracted, followed by a reconstruction stage (decoding blocks), receiving the information lost during compression from the encoder via symmetric residual connections between blocks (skip connections).

We now consider separately the introduction of the time feature (embedding) $e \in \mathbb{R}^{128}$, into the network, obtained by passing the scalar value of the current diffusion time step t through a multilayer perceptron (MLP) (Fig. 2).

It consists of two linear fully connected layers with a SiLU (sigmoid linear unit) activation function between them, which maps the scalar t to a representation e_t of fixed dimensionality 128. This choice is motivated by a balance between computational efficiency and the quality of the time representation, which is necessary for reliable encoding of the degree of image restoration. Each convolutional block (ConvBlock) includes an additional linear layer that takes e_t as input in order to map it to a space determined by the number of channels of this layer. Formally, the mapping $e_t \to \mathbb{R}^C$ is performed. The resulting vector w is expanded to a tensor of size $C \times 1 \times 1$ and is additively incorporated into the activations after the first convolutional layer. This approach makes it possible to achieve consistency of dimensionalities and to integrate the time information into the image restoration process, which is one of the most important aspects of this dynamical process.

The output of the model is the predicted clean image x_0. For each step t of the reverse process, the model sequentially processes the noisy intermediate states x_t, taking into account the conditional image I. During training, the original clean image x_0 is given and treated as the reference: the model parameters θ are optimized by minimizing the loss function defined as the mean squared error between the predicted and true samples

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,t}\Big[x_0 - \hat{x}_0(x_t, t, I)^2\Big].$$

This loss function makes it possible to directly assess the quality of image restoration, and its minimization leads to a reduction in the discrepancy between the predicted and true objects.

5 Experimental Results

The model considered in this work is a special case of diffusion probabilistic models. The experimental results presented in this section were obtained on the CIFAR-10 dataset, which contains 60000 images of size 32×32 pixels [20]. To corrupt them, additive Gaussian noise controlled by a target signal-to-noise ratio (SNR) was used. Two noise levels specified by values of 20 dB and 10 dB are considered, which is due to the need to analyze the performance of the model under both moderate and strong distortion of objects.

The noise corruption procedure is carried out as follows. For each original (clean) image, the average signal power is computed, defined as the expectation of the squared pixel values of the image

$$P_{\text{signal}} = \mathbb{E}\Big[x^2\Big].$$

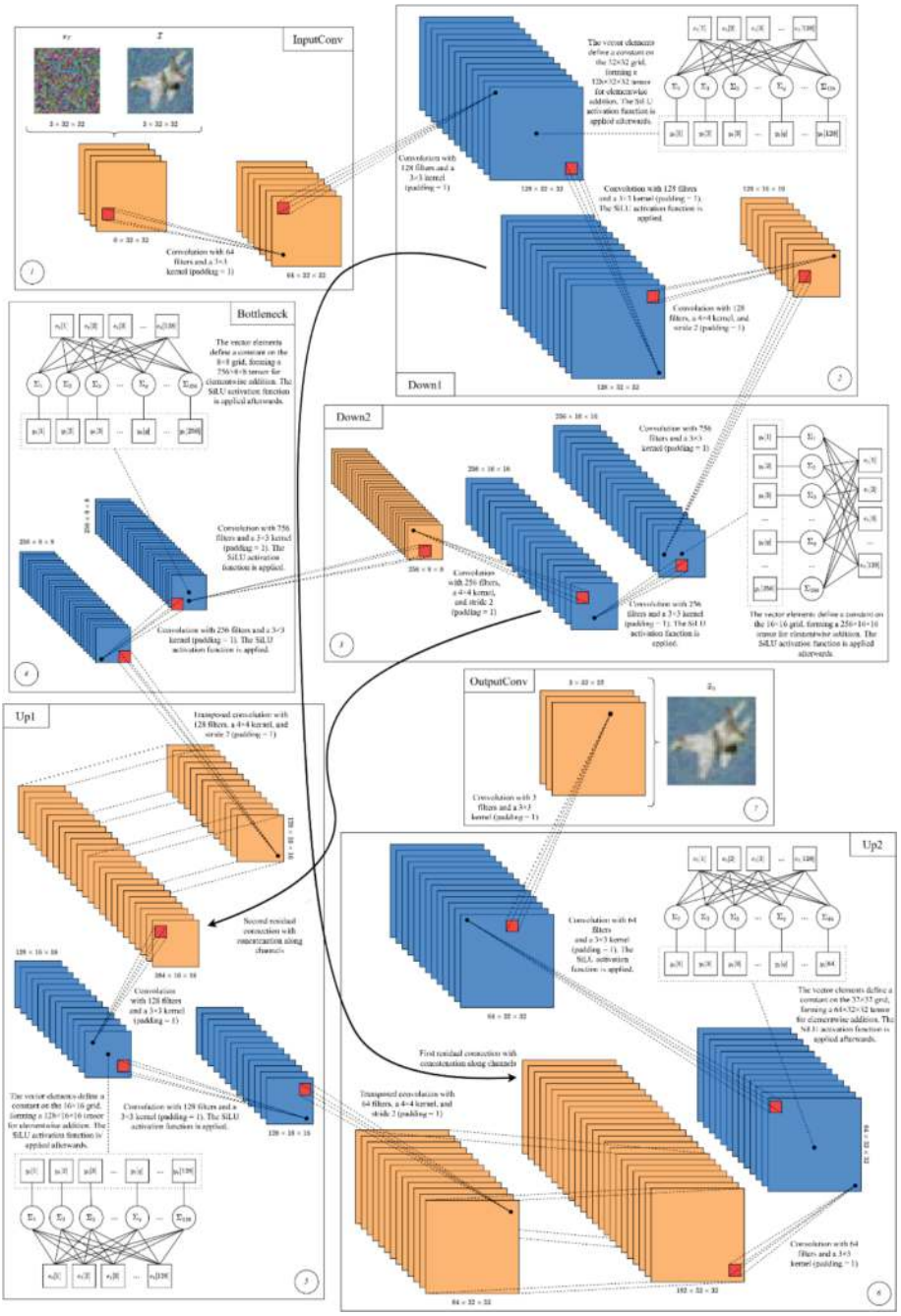


Fig. 1. Architecture of the cdDPM.

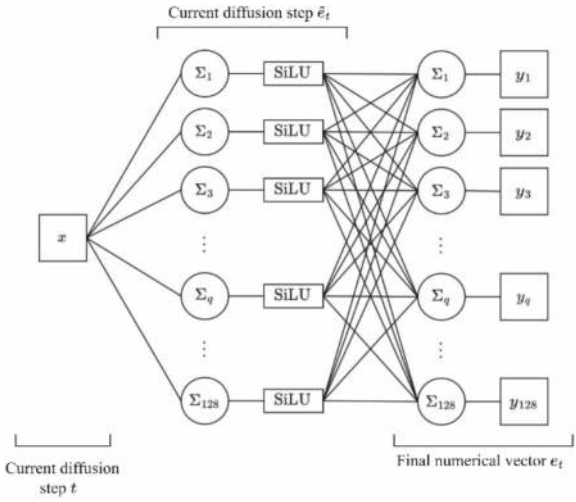


Fig. 2. Architecture of the used MLP.

Given the specified signal-to-noise ratio, the noise power is calculated as

$$P_{\text{noise}} = \frac{P_{\text{signal}}}{10^{\frac{SNR}{10}}}.$$

Based on this value, the noise is defined as a normally distributed random variable with zero mean and standard deviation

$$\sigma_{\text{noise}} = \sqrt{P_{\text{noise}}},\ \varepsilon \sim \mathcal{N}(0, \sigma_{\text{noise}}).$$

The final noisy image is formed by adding the computed noise value to the original object, followed by restricting the range of pixel values. After training the models with different numbers of epochs and diffusion process steps, the results were collected (Tables 1 and 2), and examples of the model performance were obtained (Figs. 3 and 4).

To obtain a comparative assessment of the efficiency of the proposed model, several image restoration methods are considered: the median filter, the bilateral filter, a method based on the discrete wavelet transform, and a method based on the discrete cosine transform (DCT) (Tables 3 and 4). These approaches were selected due to their widespread use in solving problems of this type, as well as their fundamentally different noise removal mechanisms, which provides a more comprehensive evaluation of the model proposed in this work. A comparison with a deep convolutional neural network (DnCNN) is also presented. The architecture implemented in the DnCNN model considered here was first described in the work [10]. It is a deep convolutional neural network with residual learning for the task of removing noise from an input noisy image. The model consists of layers: an initial convolutional layer with a convolution kernel and a ReLU activation function, a sequence of intermediate blocks including convolution without bias, batch normalization, and a nonlinear transformation (ReLU), and a final layer performing a convolution that restores the number of channels to the original one.

Table 1. Comparison of model performance across varying training epochs and diffusion-process steps at a noise level 20 dB.

Epochs	Reverse process steps	PSNR (dB)	SSIM
20	25	29,19	0,9328
	50	28,33	0,9256
	100	27,16	0,9114
40	25	30,24	0,9385
	50	29,49	0,9364
	100	29,47	0,9331
80	25	30,19	0,9369
	50	30,46	**0,9463**
	100	**30,49**	0,9409

Table 2. Comparison of model performance across varying training epochs and diffusion-process steps at a noise level 10 dB.

Epochs	Reverse process steps	PSNR (dB)	SSIM
20	25	23,39	0,7977
	50	23,26	0,7999
	100	**23,51**	0,8048
40	25	23,13	0,7896
	50	23,52	0,8022
	100	22,90	0,7727
80	25	22,87	**0,8118**
	50	22,73	0,8101
	100	22,01	0,7715

The model approximates the magnitude of the added noise, allowing the restored image to be computed as the difference between the input signal and the estimated noise component, which, from a mathematical point of view, corresponds to minimizing the mean squared error between the true signal and the obtained result.

Table 3. Comparison of the obtained result with alternative approaches at a noise level 20 dB.

Processing method	Parameters	PSNR (dB)	SSIM
Median filter	Window 3×3	25,89	0,8547
Bilateral filter	$d = 3, \sigma_{\text{color}} = 3, \sigma_{\text{space}} = 3$	27,12	0,8928

(*continued*)

Table 3. (*continued*)

Processing method	Parameters	PSNR (dB)	SSIM
Wavelet transform (db4)	lvl = 3, threshold = 0,7	27,74	0,8892
DCT	threshold = 0,1	26,33	0,8565
DnCNN	epoch = 40	**30,76**	0,9455
cdDPM	epoch = 80, $T = 50$	30,46	**0,9463**

Table 4. Comparison of the obtained result with alternative approaches at a noise level 10 dB.

Processing method	Parameters	PSNR (dB)	SSIM
Median filter	Window 3×3	21,31	0,6811
Bilateral filter	$d = 5$, $\sigma_{\text{color}} = 25$, $\sigma_{\text{space}} = 25$	21,89	0,7005
Wavelet transform (db4)	lvl = 2, threshold = 0,7	19,13	0,6058
DCT	threshold = 0,5	19,91	0,5675
DnCNN	epoch = 40	**24,75**	**0,8335**
cdDPM	epoch = 20, $T = 100$	22,87	0,8118

6 Conclusion

The experimental results confirm that the conditional diffusion probabilistic model presented in this work successfully addresses the problem of restoring images corrupted by additive Gaussian noise, demonstrating quality comparable to existing methods. In terms of PSNR and SSIM, the proposed approach achieved results on par with DnCNN, which indicates the competitiveness of cdDPM with respect to minimal neural network methods, and it also surpassed several classical filtering techniques while eliminating the need for expert parameter tuning. Unlike traditional methods focused on point estimation, the proposed approach makes it possible to determine the distribution of all possible solutions, which removes the need for manual control, since the a priori properties of images are extracted from the training data.

Although the reported results were obtained at a comparatively low image resolution 32×32 pixels and under synthetic Gaussian noise, the method has already demonstrated a number of potential advantages at this stage. First, cdDPM directly models the posterior distribution, which theoretically allows one to generate multiple plausible objects while taking into account the uncertainty of the inverse problem solution, which distinguishes it from deterministic approaches that provide a single solution. Second, the model is characterized by both a convenient modular structure and versatility: the solution can be readily extended by adding new blocks or adapting it to other types of distortions.

In this regard, further research in this direction appears to be relevant, and promising steps may include scaling the model to high-resolution images and applying cdDPM to

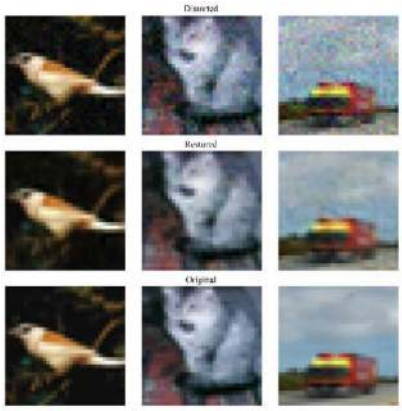


Fig. 3. Example of the model's output at a noise level 20 dB

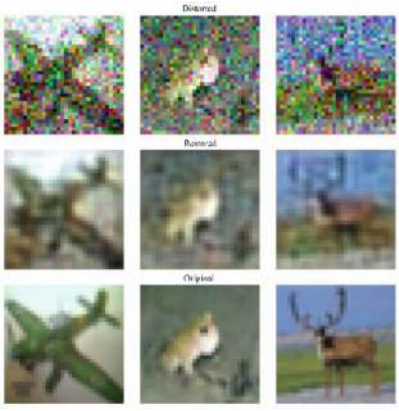


Fig. 4. Example of the model's output at a noise level 10 dB

more complex restoration scenarios. For example, suppressing real noise, characterized by a complex distribution law, or removing artifacts. Another interesting line of work is the optimization of the reverse process by incorporating existing studies in this area in order to reduce computational complexity without a significant loss of quality.

References

1. Fan, L., Zhang, F., Fan, H., Zhang, C.: Brief review of image denoising techniques. Vis. Comput. Ind. Biomed. **7**, Art. 2 (2019)
2. Goyal, B., Dogra, A., Agrawal, S., Sohi, B.S., Sharma, A.: Image denoising review: from classical to state-of-the-art approaches. Inf. Fusion. **55**, 220–244 (2020)
3. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D: Nonlinear Phenom. **60**(1–4), 259–268 (1992)

4. Buades, A., Coll, B., Morel, J.-M.: A non-local algorithm for image denoising. In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 60–65 (2005)
5. Willoughby, R.A.: Solutions of ill-posed problems (AN Tikhonov and V.Y. Arsenin). SIAM Rev. **21**(2), 266–267 (1979)
6. Elad, M., Kawar, B., Vaksman, G.: Image Denoising: the deep learning revolution and beyond—a survey paper. SIAM J. Imaging Sci. **16**(3), 1594–1654 (2023)
7. Wiener, N.: Extrapolation, Interpolation, and Smoothing of Stationary Time Series. Wiley, New York (1949)
8. Donoho, D.L.: De-noising by soft-thresholding. IEEE Trans. Inf. Theory. **41**(3), 613–627 (1995)
9. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
10. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
11. Zhang, K., Zuo, W., Zhang, L.: FFDNet: toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inf. Proces. Syst. **33**, 6840–6851 (2020)
13. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Adv. Neural Inf. Proces. Syst. **34**, 8780–8794 (2021)
14. Chan, S.H.: Tutorial on Diffusion Models for Imaging and Vision. Now Publishers, Boston (2025)
15. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: ILVR: conditioning method for denoising diffusion probabilistic models. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14347–14356 (2021)
16. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4713–4726 (2023)
17. Yang, B. et al.: Paint by example: exemplar-based image editing with diffusion models. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18381–18391 (2023)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015)
19. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of 32nd International Conference on Machine Learning (ICML 2015), pp. 2256–2265 (2015)
20. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images: Technical Report. University of Toronto, Toronto (2009)

Optimizing Alt Text Design through Human-LLM Collaboration

Yekaterina Kosova(✉) and Kirill Redkokosh

V.I. Vernadsky Crimean Federal University, Prospekt Vernadskogo 4, Simferopol 295007, Russia
lynx99@inbox.ru

Abstract. This study evaluates the efficacy of large language models (LLMs) in generating compliant alt texts for digital accessibility, focusing on their ability to balance brevity, accuracy, and WCAG 2.2 standards. Using three LLM-based chatbots (Copilot, Gemini, YandexGPT), 108 English-language descriptions were generated for 12 images (6 photographs, 6 logos) under varying length constraints (unrestricted, ≤250, ≤125 characters). The descriptions were assessed via expert-rated 5-point Likert scales for usefulness, reliability, completeness, accuracy, and literacy, alongside phantasm analysis (semantic deviations from original keywords). Key findings reveal that stringent length limits (≤125 characters) significantly degrade completeness and usefulness, while literacy remains consistently high. Notably, the quality of generated descriptions exceeded median benchmarks (average score: 4.2/5.0), confirming the methodology's effectiveness in balancing compliance and creativity. Phantasms–observed in 38.9% of outputs–correlate with creative liberties in unrestricted texts, particularly for Copilot. Regression models highlight 125-character limits as the strongest negative predictor, while image type (logos vs. photographs) improves reliability due to structural predictability. LLMs exhibit divergent performance: YandexGPT adapts better to constraints, whereas Copilot prioritizes precision over speed. The study underscores the necessity of hybrid workflows, combining human expertise for context and LLMs for scalability, to mitigate digital exclusion.

Keywords: Digital accessibility · Large language models · Alt text generation · WCAG compliance · Semantic distortions · Human-AI collaboration

1 Introduction

Digital accessibility, defined as the ability of electronic environments to adapt to the needs of users with diverse limitations [1, 2], is a critical requirement for over a quarter of a billion people with visual impairments [3]–one of the most vulnerable groups to digital exclusion [4, 5]. According to an analysis of one million websites, more than one-fifth of images on homepage lack alternative text (alt text) [6], exacerbating digital inequality for users with low vision [7]. The Web Content Accessibility Guidelines (WCAG) 2.2 [8] mandate that all non-text content — including images, charts, and interactive elements—must have text alternatives to ensure equitable access. Success Criterion 1.1.1

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 251–262, 2027.
https://doi.org/10.1007/978-3-032-27834-0_24

("Non-text Content"), classified as Level A conformance (the highest priority), explicitly requires text descriptions to convey the purpose, function, or meaning of visual content. Failure to comply excludes users who rely on assistive technologies, such as screen readers or Braille displays, from accessing critical information. Concurrently, the widespread increase in web content graphical complexity [6] necessitates the large-scale implementation of text alternatives to comply with WCAG 2.2 standards [8].

Existing automation methods for generating text alternatives, including computer vision and neural network algorithms, face limitations in accuracy, ethical compliance, and contextual relevance [9–15]. Manual creation of alternative texts is also problematic due to authors' insufficient expertise [15, 16], highlighting the need for optimized solutions.

Advances in large language models (LLMs) offer promising opportunities for combining human creativity with machine processing. LLMs can generate coherent texts, analyze semantics, and adapt writing styles [17–20], positioning them as ideal "digital assistants" for content creators. However, the effectiveness of LLMs in developing image-related text descriptions remains underexplored.

Research objective: to test the hypothesis that a collaborative human-LLM methodology, based on predefined keywords, enables the creation of image-relevant alternative text descriptions.

In accordance with the research aim, the following research questions have been formulated:

1. Can LLMs overcome the visual-contextual gap to generate pragmatically useful, semantically exhaustive, and factually reliable image descriptions solely through keyword-based input?
2. Through what approaches can the quality of textual descriptions produced by a LLM based on given keywords be evaluated? What aspects might influence this quality?
3. What recommendations should be provided to authors of alternative textual descriptions to ensure the correct application of LLM tools in their professional activities?

2 Material and Methods

The study was conducted from April to August 2024, encompassing study design development, image selection, description generation, description quality assessment, and data analysis. Three chatbots were employed: Copilot (Microsoft®; launched 2023/02/07), Gemini (Google®; launched 2023/03/21), and YandexGPT (Yandex®; launched 2023/05/17). Twelve images (six story photographs and six text-containing logos) were analyzed.

Descriptions were generated via a five-step algorithm: (1) image examination; (2) keyword list creation (10–20 items) with content-specific parameters (covering color, location, background, foreground, and character details like appearance, gender, age, placement, emotions, and behavior for photos; font style, graphic objects, and element relationships for logos); (3) formulation of three prompts ("Create a text description using keywords: …"; "Shorten to 250 characters with spaces, preserving meaning"; "Shorten to 125 characters with spaces, preserving meaning"); (4) dataset recording (fields: chatbot, image number, type (photo/logo), keywords, keywords' count, keywords' length,

keywords' generation time (144–510 s, timed by stopwatch), description text, required and actual lengths, and generation time); (5) browser data clearance. As a result, 108 English-language descriptions were generated and added to the dataset.

Two experts evaluated descriptions using five 5-point Likert scales. These included *usefulness*, measuring how applicable descriptions are for individuals unable to view images (1 = absolutely useless, 5 = extremely useful); *reliability*, assessing alignment with factual accuracy (1 = completely unreliable or fabricated, 5 = fully truthful or perceived as such); *completeness*, gauging sufficiency for grasping the image's essence (1 = entirely insufficient for basic understanding, 5 = thoroughly comprehensive); *accuracy*, evaluating clarity in conveying visual and contextual details (1 = highly imprecise, 5 = exceptionally precise); and *literacy*, focusing on adherence to grammatical and linguistic standards (1 = severely error-ridden, 5 = flawlessly composed). Both experts pre-tested the method on 15 descriptions.

The generated descriptions were also checked for phantasms – text elements containing semantic deviations from the original keywords: liberal interpretations, false assertions, or intrusive narratives that distort objective perception of visual content.

Statistical analysis utilized Microsoft Excel® and Python, employing Pearson correlation, linear regression, ordinal regression, and unpaired two-Samples t-test (significance threshold $p\text{-}value \leq 0.05$). The results of the study, including image data, text descriptions and quality assessments, are posted in the Mendeley Data [21]. The DeepSeek chatbot (DeepSeek®; launched 2025/01/20) served as an auxiliary tool for both research data analysis and translation enhancement.

3 Results

3.1 Description Length and Generation Time

The generated text descriptions exhibited significant variability in length, ranging from 54 to 1,354 characters including spaces. Analysis revealed that chatbots tend to produce lengthy descriptions when unrestricted, yet generally adhere to length constraints specified in prompts. Out of 72 descriptions generated with such constraints, 63 (87.5%) complied with the specified parameters. Violations occurred most frequently in the 125-character limit category, highlighting the difficulty of adhering to ultra-short constraints. Statistics on the lengths of text descriptions are given in Table 1.

A comparison of individual chatbot models revealed notable disparities. YandexGPT demonstrated the lowest compliance: 29.1% of its descriptions exceeded limits, whereas Copilot and Gemini each showed only 4.2% non-compliance. This discrepancy suggests differences in how these models algorithmically process user instructions.

The range of lengths in constrained descriptions (129–280 characters for the 250-character limit; 54–201 characters for the 125-character limit) indicates uneven utilization of available "length allowances," even when limits were nominally respected. Furthermore, standard deviations across all categories underscore length instability as an indicator of chatbot predictability. The greatest variability occurred in the unrestricted descriptions ($\sigma = 244.0$), reflecting inherent generative freedom, while the constrained conditions showed the reduced deviations – though still notable (e.g., YandexGPT's $\sigma = 38.1$ under the 125-character limit) (Table 1).

Table 1. Length of English-language text descriptions generated by chatbots.

Chatbot model	Statistics	Length of text descriptions according to length limits		
		No length limit	≤250 characters	≤125 characters
Copilot	$M \pm \sigma$ (*n*)	843.9 ± 218.3 (12)	194.3 ± 22.6 (12)	107.8 ± 13.6 (12)
	Minimum	541	161	82
	Maximum	1354	242	132
Gemini	$M \pm \sigma$ (*n*)	453.7 ± 155.4 (12)	200.1 ± 46.4 (12)	93.4 ± 24.3 (12)
	Minimum	199	129	54
	Maximum	781	280	123
YandexGPT	$M \pm \sigma$ (*n*)	506.5 ± 140.6 (12)	211.3 ± 32.2 (12)	141.7 ± 38.1 (12)
	Minimum	294	149	93
	Maximum	707	260	201
Overall	$M \pm \sigma$ (*n*)	601.4 ± 244.0 (36)	201.9 ± 34.8 (36)	114.3 ± 33.5 (36)
	Minimum	199	129	54
	Maximum	1354	280	201

Notes M–arithmetic mean, σ–standard deviation, n–number of descriptions.

To identify relationships between description length and multiple factors (chatbot model, generation time, presence of fantasized elements, keyword count/length, image type), a multivariate regression analysis was applied. The results revealed two significant predictors: the fantasized elements in Copilot's texts increased description length by 12% (model coefficient $\beta = 0.12$; *p-value* $= 0.02$), suggesting a link between "creative" deviations and verbalized details in this model; generation time positively correlated with length ($\beta = 0.28$; *p-value* $= 0.04$) across all chatbots, implying longer processing consistently yields more verbose outputs. Other variables, including keyword parameters and image type, showed no statistical significance.

The temporal dynamics of description generation revealed operational disparities among chatbots. Processing times fluctuated across the dataset, ranging from rapid 2-s responses to computationally intensive 43-s intervals ($M = 7.52$, $\sigma = 8.53$). A pronounced efficiency hierarchy emerged: Copilot demonstrated markedly prolonged processing durations ($M = 15.86$, $\sigma = 10.55$), contrasting sharply with Gemini's ($M = 3.50$, $\sigma = 1.06$) and YandexGPT's ($M = 3.19$, $\sigma = 1.67$) optimized workflows.

Multivariate regression modeling exposed relationships between temporal expenditure and operational parameters. Four factors demonstrated statistical significance in reducing generation latency: YandexGPT ($\beta = -12.67$, *p-value* < 0.001) and Gemini ($\beta = -12.36$, *p-value* < 0.001) usage; 125-character limits ($\beta = -7.31$, *p-value* < 0.001) and 250-character limits ($\beta = -6.06$, *p-value* < 0.001). Notably, visual content typology (photographs vs. logos) showed null predictive power (*p-value* > 0.05).

While stringent length constraints accelerated text generation systematically, Copilot's paradoxically prolonged processing times may stem from its multi-layered semantic verification protocols – a quality-control mechanism that prioritizes precision over speed.

These temporal patterns provide crucial insights for optimizing human-AI collaboration workflows, particularly in time-sensitive applications requiring rapid yet precise textual generation.

3.2 Phantasmic Distortions in Generated Texts

The analysis revealed phantasms in 38.9% of generated texts (42 cases), with their distribution reflecting dependencies on chatbot model architectures and operational conditions. As illustrated in Fig. 1, the synthesis of alternative realities intensifies in the absence of strict length constraints, a pattern consistent across all chatbots regardless of image type.

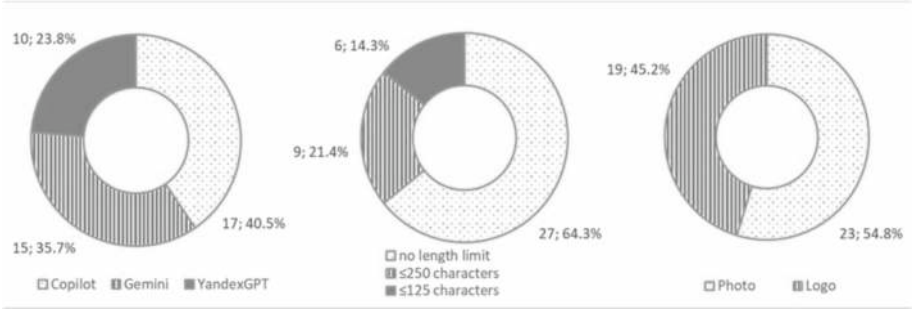


Fig. 1. The number of English-language text descriptions with phantasms, $N = 42$. On the left is the frequency distribution according to the chatbot model, in the center–depending on the required length of descriptions, on the right–by image type.

The proposed taxonomy of phantasmic artifacts, developed by the authors, categorizes these distortions into five interconnected types: *ornamental excesses* – epithets and metaphors overloading description; *speculative constructs* – subjective inferences projecting hidden meanings; *emotional projections* – attribution of nonexistent affective states to images; *semantic noise* – vacuous clichés masquerading as analysis; and *factological mystifications* – false (distorted or invented) misrepresentations of visual reality.

The distribution frequencies of phantasms and examples of their appearance in text descriptions are given in Table 2.

The analysis uncovered the layered coexistence of phantasmic elements – ranging from singular deviations (23 instances, 54.76% of anomalous texts) to intricate semantic collages combining four distortion types (3 cases, 7.14%). Notably, the majority (7 out of 9) of descriptions featuring three or four phantasm types originated from Copilot. This imbalance highlights a fundamental distinction in the chatbot's architectural approach to processing creative constraints, underscoring its propensity for semantic complexity compared to other models.

Table 2. Occurrence frequencies and examples of different types of phantasms in English-language text descriptions.

Phantasm type; frequency of occurrence (%)	Example of phantasm in description*	Notes	Image title (Author, Year)
Ornamental excesses; 20 (18.5)	… *majestic* spruce tree... creating a *heartwarming* scene of *childhood delight* and *winter magic*... It's a *timeless moment*, *frozen* in black and white…	epithets and metaphors; not in keywords	"Snow Shower" (Borodulin 1960)
Speculative constructs; 28 (25.9)	… The symmetry and discipline in the formation are a *testament to their training and dedication*...	value judgment; not in keywords	"Olympics-80" (Abaza 1980)
Emotional projections; 11 (10.2)	… freeze-frame of a moment of *pure, unadulterated happiness*, a snapshot of *childhood delight* in the heart of nature…	emotional experiences; not in keywords	"Full Speed" (Drysdale 1969)
Semantic noise; 4 (3.7)	… This *description effectively paints a picture with words, highlighting the key elements and details*…	redundant information; not in keywords	Nestlé S.A. Logo (Nestlé 2023)
Factological mystifications; 10 (9.3)	… A child in a fur coat *kneels* in a snowy desert...	false information; not in keywords	"Teddy Bear" (Romanov 2020)

Note *–phantasms of the corresponding type are highlighted in italics.

3.3 Quality of the Generated Texts

Expert evaluation of the generated text descriptions revealed consistent outperformance against median benchmarks (Table 3). High ratings predominated: 4-point scores accounted for 37.8% of cases (204 descriptions), while top-tier 5-point evaluations comprised 34.1% (184). Critically low scores (1–2 points) represented a marginal 6.5% (35). Literacy quality peaked at 4.1–4.9 points, whereas functional metrics displayed variance: accuracy (2.8–4.1), reliability (4.3–4.6), usefulness (2.8–4.4), and semantic completeness (2.5–4.8).

Descriptive statistics suggest the following patterns: usefulness and completeness sharply decrease under stricter length constraints, while accuracy moderately degrades

with text shortening. Reliability and literacy, however, remain stable regardless of required description length. Copilot and Gemini exhibit critical vulnerability to restrictions, whereas YandexGPT demonstrates adaptability to limitations. Overall, length constraints create a trade-off between detail and brevity, particularly critical for completeness and usefulness. All chatbots maintain high literacy scores ($\geq$4.1) even under stringent constraints, underscoring the maturity of neural network algorithms in adhering to linguistic norms.

Table 3. Quality of English-language text descriptions generated by chatbots.

Quality scale	Chatbot model	Quality rating, $M \pm \sigma(n)$		
		No length limit	$\leq$250 characters	$\leq$125 characters
Usefulness	Copilot	4.33 ± 0.78 (12)	4.17 ± 0.94 (12)	3.33 ± 1.07 (12)
	Gemini	4.08 ± 0.79 (12)	3.75 ± 0.45 (12)	2.83 ± 0.39 (12)
	YandexGPT	4.42 ± 0.67 (12)	3.92 ± 0.67 (12)	3.17 ± 0.72 (12)
	Overall	*4.28 ± 0.74 (36)*	*3.94 ± 0.71 (36)*	*3.11 ± 0.78 (36)*
Reliability	Copilot	4.25 ± 0.75 (12)	4.58 ± 0.51 (12)	4.42 ± 0.79 (12)
	Gemini	4.33 ± 0.65 (12)	4.42 ± 0.67 (12)	4.33 ± 0.78 (12)
	YandexGPT	4.33 ± 0.49 (12)	4.42 ± 0.67 (12)	4.42 ± 0.67 (12)
	Overall	*4.31 ± 0.62 (36)*	*4.47 ± 0.61 (36)*	*4.39 ± 0.73 (36)*
Completeness	Copilot	4.83 ± 0.39 (12)	3.75 ± 0.75 (12)	2.58 ± 1.00 (12)
	Gemini	4.25 ± 0.62 (12)	3.50 ± 0.52 (12)	2.50 ± 0.52 (12)
	YandexGPT	4.50 ± 0.52 (12)	3.33 ± 0.78 (12)	2.67 ± 0.78 (12)
	Overall	*4.53 ± 0.56 (36)*	*3.53 ± 0.70 (36)*	*2.58 ± 0.77 (36)*
Accuracy	Copilot	4.08 ± 0.67 (12)	3.92 ± 0.67 (12)	3.58 ± 0.79 (12)
	Gemini	3.58 ± 0.79 (12)	3.50 ± 0.52 (12)	2.83 ± 0.39 (12)
	YandexGPT	4.08 ± 0.79 (12)	3.67 ± 0.78 (12)	3.08 ± 0.79 (12)
	Overall	*3.92 ± 0.77 (36)*	*3.69 ± 0.67 (36)*	*3.17 ± 0.74 (36)*
Literacy	Copilot	4.83 ± 0.39 (12)	4.83 ± 0.39 (12)	4.50 ± 0.52 (12)
	Gemini	4.58 ± 0.51 (12)	4.50 ± 0.67 (12)	4.08 ± 0.67 (12)
	YandexGPT	4.75 ± 0.45 (12)	4.92 ± 0.29 (12)	4.83 ± 0.39 (12)
	Overall	*4.72 ± 0.45 (36)*	*4.75 ± 0.50 (36)*	*4.47 ± 0.61 (36)*

Notes M–arithmetic mean, σ–standard deviation, n–number of descriptions.

A multilayer ordinal regression analysis was employed to verify multifactorial dependencies, incorporating variables: actual/required length parameters, chatbot model, number and length of keywords, generation time, image type, and presence of phantasmic artifacts. Resulting patterns (Table 3) confirm selective factor influence on specific metrics. The exception–literacy–defied modeling as no statistically viable predictors

emerged, highlighting the universal language competence of models across operational environments.

The results presented in Table 4 can be interpreted as follows:

- *usefulness* declined under 125-character length constraints and with increased text generation time. The use of Gemini and YandexGPT chatbots also negatively impacted usefulness, indicating their limited applicability for users unable to view images;
- *reliability* decreased when using Gemini and YandexGPT, as well as with longer keyword sequences. However, descriptions for logos exhibited higher reliability than those for photographs, likely due to logos' structural predictability (brand colors, geometry), which reduces interpretation errors;
- *completeness* was compromised under 125-character limits and when using lengthy keyword lists. Gemini and YandexGPT produced less comprehensive descriptions compared to Copilot;
- *accuracy* diminished in length-constrained descriptions and when employing Gemini and YandexGPT.

Overall, the most significant negative impact on quality stemmed from 125-character constraints, confirming that ultra-concise formats critically disrupt informational integrity. For all analyzed parameters, actual description length, presence of phantasms, and Copilot usage were not significant predictors of quality improvement or degradation. AIC metrics revealed that models for completeness (AIC = 205.9) and reliability (AIC = 203.8) demonstrated superior explanatory power compared to accuracy (AIC = 229.7) and, particularly, usefulness (AIC = 238.8).

To identify structural interdependencies between text description quality metrics, a correlation mapping methodology was employed (Table 5). The Pearson correlation matrix visualized hidden patterns of metric synergy, revealing coherent alignments in language model outputs.

Analysis of the correlation matrix using the Chaddock scale revealed multi-tiered relationships between quality parameters. The positive correlation between usefulness and completeness ($r = 0.75$), as well as usefulness and accuracy ($r = 0.70$), aligns with the "strong" category (0.7–0.9), indicating tight interdependence of pragmatic text characteristics. Links between reliability and accuracy ($r = 0.38$) and completeness and accuracy ($r = 0.59$) fall into the "weak" (0.3–0.5) and "moderate" (0.5–0.7) strength ranges, though the insignificance of some coefficients ($p\text{-}value > 0.05$) underscores the conditional nature of these interactions. Literacy exhibits statistical isolation: its correlations with other parameters remain within the "negligible" association threshold ($r < 0.3$), while its negative association with reliability ($r = -0.14$) resides in the statistical noise zone. Thus, the Chaddock scale confirms the dominance of the "usefulness–completeness–accuracy" cluster as the qualitative core, whereas literacy operate independently, reflecting divergent priorities between linguistic-technical and content-driven evaluation.

Table 4. Assessing the influence of independent variables on the quality of English-language text descriptions.

Quality scale	Significant variables	Coefficient (t)	Residual deviance (AIC)
Usefulness	Generation time	−5.79 (−2.37)	210.8 (236.8)
	Required length (≤125 characters)	−3.71 (−3.70)	
	Gemini Chatbot	−2.88 (−3.81)	
	YandexGPT Chatbot	−2.15 (−2.72)	
Reliability	Keywords' length	−2.59 (−1.75)	179.8 (203.8)
	Image type – logo	1.21 (2.36)	
	YandexGPT Chatbot	−1.01 (−1.31)	
	Gemini Chatbot	−0.99 (−1.36)	
Completeness	Required length (≤125 characters)	−5.19 (−4.20)	177.9 (205.9)
	Keywords' length	−4.25 (−2.76)	
	Gemini Chatbot	−2.03 (−2.60)	
	YandexGPT Chatbot	−1.92 (−2.25)	
Accuracy	Required length (≤125 characters)	−2.67 (−2.69)	203.7 (229.7)
	Gemini Chatbot	−2.00 (−2.92)	
	Required length (≤250 characters)	−0.96 (−1.15)	
	YandexGPT Chatbot	−0.94 (−1.33)	

Notes t–t-criterion for ordinal regression, the four largest t by absolute value are included in the model; AIC–Akaike information criterion.

Table 5. Pearson correlation matrix of English-language text descriptions quality parameters.

	Usefulness	Reliability	Completeness	Accuracy	Literacy
Usefulness	1.00	0.50	0.75	0.70	0.01
Reliability	0.50	1.00	0.25*	0.38	−0.14
Completeness	0.75	0.25*	1.00	0.59*	0.16*
Accuracy	0.70	0.38	0.59*	1.00	0.19
Literacy	0.01	−0.14	0.16*	0.19	1.00

Note *–the correlation is insignificant (p-*value* > 0.05).

4 Discussion

This study evaluated the effectiveness of LLMs in generating English-language alternative texts for images. The results demonstrated that LLMs can produce descriptions with high average quality ($M = 4.2/5.0$) in compliance with WCAG 2.2 requirements while maintaining a balance between brevity and informativeness. These findings are comparable to the outcomes of a similar study conducted on Russian-language alt texts for same images [22], yet demonstrate more robust patterns in the behavior of individual chatbots. A comparative analysis of quality ratings for English- and Russian-language descriptions across all metrics and chatbots further showed statistical similarity between the samples ($p\text{-}value > 0.5$), suggesting that the language of description does not significantly influence the measured quality parameters.

The chatbots successfully transformed predefined keywords into coherent texts, adapting to length constraints. However, reducing descriptions to 125 characters diminished completeness, usefulness, and accuracy. This aligns with findings by [23], who notes that ultra-concise formats disrupt informational integrity yet remain essential for screen reader compatibility. At the same time, longer alternative descriptions – which scored higher on semantic quality metrics in this study – are applicable in web documents as image captions or extended descriptions in linked files [10, 23].

A key finding was the identification of phantasms, detected in 38.9% of descriptions. Unlike AI "hallucinations" [24] or "bullshit" [25], elements such as ornamental excesses and speculative constructs often serve literary embellishment but require scrutiny due to misinformation risks (9.3% of phantasm cases). This underscores the need for post-generation human verification, as advocated by [23].

LLMs significantly accelerate the workflow: the combined time for keyword creation and text generation (5.5 min per image) is one order of magnitude shorter than manual creation (up to 60 min for complex images [23]), reducing cognitive load for authors, especially those working in non-native languages. These results corroborate [11], who highlight author fatigue during prolonged manual description tasks.

Regression analysis revealed no significant negative impact of Copilot on quality metrics (usefulness, reliability, completeness, accuracy), unlike Gemini and YandexGPT, whose use statistically degraded these parameters. Copilot's stability likely stems from its architecture, which mitigates semantic distortions even under strict constraints.

Correlation analysis confirmed interconnections among quality metrics: strong links between accuracy, usefulness, and completeness ($r = 0.59$–0.75) indicate their interdependence, while literacy (consistently high-scoring) remained autonomous ($r < 0.2$), reflecting the maturity of linguistic algorithms.

Based on these findings, we propose the following recommendations for alt text authors: (1) use concise keywords emphasizing critical image elements; (2) clear browser cache to minimize contextual distortions; (3) mandatorily verify texts for disinformative phantasms.

This study lays the groundwork for standardized hybrid protocols combining LLM creativity with expert oversight, crucial for reducing digital inequality.

Limitations include the lack of evaluation by users with visual loss and comparative analysis with manually crafted descriptions. Future work should integrate LLMs into

tools with automated phantasm detection and expand research to other languages and content types (e.g., infographics).

5 Conclusion

The systematic use of text descriptions for visual content reduces the risk of digital inequality for users with profound visual impairments. Optimizing the development of alternative texts requires hybrid methods that combine artificial intelligence techniques and human expertise. The proposed methodology for creating text descriptions using LLM-based chatbots as digital assistants enables the generation of relevant alternatives for static images, characterized by high literacy, sufficient completeness, accuracy, reliability, and usefulness for individuals unable to perceive visual content. The study identified predictors of quality in generated descriptions and evaluated the semantic characteristics of the final texts, allowing for the formulation of practical recommendations for content creators on integrating LLMs into their workflows.

Future research directions include embedding LLMs into tools for automated analysis of semantic distortions and extending studies to other languages and content types. This will facilitate the creation of standardized protocols that balance the speed of automated generation with the precision of manual oversight.

References

1. Kulkarni, M.: Digital accessibility: challenges and opportunities. IIMB Manag. Rev. **31**(1), 91–98 (2019). https://doi.org/10.1016/j.iimb.2018.05.009
2. Chemnad, K., Othman, A.: Digital accessibility in the era of artificial intelligence – bibliometric analysis and systematic review. Front. Artif. Intell. **7**, 1349668 (2024). https://doi.org/10.3389/frai.2024.1349668
3. World Blind Union. https://worldblindunion.org/. Last Accessed 15 Sep 2025
4. Kerdar, S.H., Bächler, L., Kirchhoff, B.M.: The accessibility of digital technologies for people with visual impairment and blindness: a scoping review. Discov. Computing. **27**(24) (2024). https://doi.org/10.1007/s10791-024-09460-7
5. Phochai, T., Setthasuravich, P., Pukdeewut, A., Wetchakama, S.: Bridging the digital disability divide: determinants of internet use among visually impaired individuals in Thailand. Disabilities. **4**, 696–723 (2024). https://doi.org/10.3390/disabilities4030043
6. The WebAIM million: Images and alternative text. https://webaim.org/projects/million/#alttext. Last Accessed 15 Mar 2025
7. Jung, C., Mehta, S., Kulkarni, A., Zhao, Y., Kim, Y.-S.: Communicating visualizations without visuals: investigation of visualization alternative text for people with visual impairments. IEEE Trans. Vis. Comput. Graph. **28**(1), 1095–1105 (2022). https://doi.org/10.1109/TVCG.2021.3114846
8. Web Content Accessibility Guidelines (WCAG) 2.2. https://www.w3.org/TR/WCAG22/. Last Accessed 15 Sep 2025
9. Sharma, H., Agrahari, M., Singh, S.K., Firoj, M., Mishra, R.K.: Image captioning: a comprehensive survey. In: In: 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), pp. 325–328. IEEE, Mathura (2020). https://doi.org/10.1109/PARC49193.2020.236619

10. Hanley, M., Barocas, S., Levy, K., Azenkot, S., Nissenbaum, H.: Computer vision and conflicting values: describing people with automated alt text. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), pp. 543–554. ACM, New York (2021). https://doi.org/10.1145/3461702.3462620
11. Mack, K., Cutrell, E., Lee, B., Morris, M.R.: Designing tools for high-quality alt text authoring. In: Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21), pp. 1–14. ACM, New York (2021). https://doi.org/10.1145/3441852.3471207
12. Jeong, H., Chun, V., Lee, H., Oh, S.Y., Jung, H.: WATAA: web alternative text authoring assistant for improving web content accessibility. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion), pp. 41–45. ACM, New York (2023). https://doi.org/10.1145/3581754.3584127
13. Chang, R.-C., Liu, Y., Guo, A.: WorldScribe: towards context-aware live visual descriptions. In: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), pp. 1–18. ACM, New York (2024). https://doi.org/10.1145/3654777.3676375
14. Katzman, J. et al.: Taxonomizing and measuring representational harms: a look at image tagging. In: Proceedings of the AAAI Conference on Artificial Intelligence. **37**(12), 14277–14285 (2023). https://doi.org/10.1609/aaai.v37i12.26670
15. Edwards, E.J., Gilbert, M., Blank, E., Branham, S.M.: How the alt text gets made: what roles and processes of alt text creation can teach us about inclusive imagery. ACM Transactions on Accessible Computing. **16**(2), 1–18 (2023). https://doi.org/10.1145/3587469
16. Chintalapati, S.S., Bragg, J., Wang, L.L.: A dataset of alt texts from HCI publications: analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In: Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22), pp. 1–12. ACM, New York (2022). https://doi.org/10.1145/3517428.3544796
17. Kasneci, E. et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023). https://doi.org/10.1016/j.lindif.2023.102274
18. Myers, D. et al.: Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. Clust. Comput. **27**, 1–26 (2024). https://doi.org/10.1007/s10586-023-04203-7
19. Labadze, L., Grigolia, M., Machaidze, L.: Role of AI chatbots in education: systematic literature review. Int. J. Educ. Technol. High. Educ. **20**(56) (2023). https://doi.org/10.1186/s41239-023-00426-1
20. Raiaan, M.A.K. et al.: A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE access. **12**, 26839–26874 (2024). https://doi.org/10.1109/ACCESS.2024.3365742
21. Kosova, Y., Redkokosh, K.: Composing alt text using large language models: dataset in English. Mendeley Data. **V1** (2024). https://doi.org/10.17632/szh5zhpgxh.1
22. Kosova, Y.A., Redkokosh, K.I.: Composing alternative text for images using large language models. Modern Information Technologies and IT-Education. **20**(3) (2024) http://sitito.cs.msu.ru/index.php/SITITO/article/view/1122
23. Richard, B.: Key issues affecting the inclusion of alt text in scholarly PDF publications. Logos. **34**(1), 44–60 (2023). https://doi.org/10.1163/18784712-03104058
24. Ji, Z. et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 248 (2023). https://doi.org/10.1145/3571730
25. Hicks, M.T., Humphries, J., Slater, J.: ChatGPT is bullshit. Ethics Inf. Technol. **26**, 38 (2024). https://doi.org/10.1007/s10676-024-09775-5

Cascade Classification Algorithm for Malware Detection in Portable Executable Files Based on Static Analysis

Alexander V. Kozachok[1(✉)], Shamil G. Magomedov[1], Andrey V. Kozachok[1], and Sergei S. Matovykh[2]

[1] MIREA – Russian Technological University, Moscow, Russian Federation
{kozachok_a,magomedov_sh,kozachok}@mirea.ru
[2] Academy of Federal Security Guard Service of the Russian Federation, Oryol, Russian Federation

Abstract. This paper presents a comprehensive study on developing a cascade classification algorithm for detecting malicious code in Portable Executable (PE) files using static analysis methods. The proposed approach combines a structural model of PE files with a two-level cascade classifier architecture. The structural model includes 333 classification features organized into six groups, with a differentiated assessment of their importance. Feature space optimization using Principal Component Analysis and Isolation Forest methods allowed reducing the number of features to 40 most informative ones without significant accuracy loss. The cascade classifier employs a Decision Tree model trained on 10 features at the first level (Recall 0.990) and a Random Forest model on 40 features at the second level (F1-score 0.987). Classification thresholds were optimized using Youden's J-statistic. Experiments on a representative dataset of 34,026 PE files (17,992 malicious and 16,034 legitimate) demonstrated that the proposed cascade reduces average analysis time by 5–12% compared to a single-level model while maintaining comparable classification quality when the malicious traffic share is below 20.6%. The practical significance lies in the potential integration into antivirus gateways and endpoint protection systems requiring fast response and high detection completeness.

Keywords: Malware detection · Static analysis · Portable Executable files · Cascade classifier · Machine learning · Feature importance assessment · Youden's index

1 Introduction

The current stage of information technology development is accompanied by a significant increase in the quantity and diversity of information security threats, among which malicious software (malware) poses the greatest danger. Executable files of Windows operating systems, known as Portable Executable (PE), play

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 263–273, 2027.
https://doi.org/10.1007/978-3-032-27834-0_25

a significant role among such threats [1]. This file format is widely used for distributing viruses, trojans, ransomware, and other types of malware.

Two approaches are traditionally used for malware detection and classification: static and dynamic analysis. Static analysis methods allow examining an executable file without actually running it in the operating system [2]. The advantages of static methods include high processing speed on large datasets (no need to deploy sandboxes or virtual environments) and no risk of negative impact on the research system. However, static methods are subject to several disadvantages, the main one being vulnerability to obfuscation and packing of malicious code [3,22].

Dynamic analysis involves running and observing the behavior of a suspicious file in a virtualized environment (sandboxes, emulators, virtual machines) and analyzing program behavior: registering system calls, memory read/write operations, network activity, registry changes, etc. [4]. The advantage of this approach is high reliability in detecting malware; however, its disadvantages include significant resource consumption, reduced performance, and scalability complexity.

Despite known limitations (vulnerability to packers and obfuscation), static analysis of PE files remains in demand due to the speed and safety of research [25]. In this regard, an urgent task is to create new models that increase the efficiency of malware detection. This paper addresses the following objectives:

- Develop a structural model of PE files containing malicious code that combines heterogeneous features of a single object in a common system.
- Propose a cascade classification algorithm that reduces computational costs while maintaining detection quality.
- Conduct experimental validation of the proposed approach on a representative dataset.

2 Literature Review

Modern approaches actively use machine learning algorithms to classify Windows executable files as malicious or legitimate based on a set of extracted static features. Such features include PE format structural fields (headers, sections, import tables), bytecode statistics, strings, and other elements. In the study [5], over 150 features were extracted from PE files, including 52 PE header fields and 100 section characteristics. The use of ensemble models such as gradient boosting and Random Forest achieved detection accuracy of 99.5% with a low error rate (0.47%).

Another study [6] confirmed the effectiveness of combining heterogeneous static features (headers, imports, byte n-grams, etc.) into a single feature vector for subsequent classification, which ensured detection of previously unknown malware samples with high accuracy. The importance of optimizing the number of features is noted, as their excessive number leads to model overfitting and unnecessary computational costs [7].

One of the key directions of static analysis is the study of opcode sequences (machine instructions) obtained after disassembling PE files [8]. Opcodes can contain patterns characteristic of malware. In [9], machine learning algorithms were compared for classification tasks based on opcodes; the best results were achieved by Random Forest and Decision Tree methods with accuracy of approximately 96–98%.

The N-gram approach involves analyzing sequences of n consecutive bytes or instructions as features. In [10], an approach with opcode n-grams was applied, where sequences of commands of length 2, 3, and 4 were extracted from each file. Training an ensemble of classifiers on these features achieved high accuracy up to 97% for some families [11]. A comprehensive survey of machine learning techniques for malware analysis [24] confirms the effectiveness of such approaches.

In recent years, deep learning methods have gained popularity, allowing automatic extraction of complex features from PE file content [12,23]. An interesting method is PE file visualization – converting code into an image (grayscale). In [13], binary code was converted to images, and a convolutional neural network was trained to recognize similarity between "images" of malware programs from the same family, achieving accuracy of approximately 95%.

Significant attention is paid to analyzing structural metadata of PE files such as DOS Header, File Header, Optional Header, import and export tables, and resources [14]. The study [15] implemented an ensemble classifier (ExtraTrees) trained on 68 numerical features from the PE file header and achieved 98.7% accuracy with 1.3% error.

The concept of cascade classifiers was first proposed in the context of image processing by Viola and Jones [16]. In the field of information security, cascade model concepts have been developed in works [17]. The authors developed architectures including several levels of PE file filtering using clustering and neural networks.

One of the current tasks in building classifiers for malware detection is selecting the threshold value that separates positive (malicious) and negative (legitimate) classes. In several works [18,19], Youden's J-statistic is proposed as a rational criterion for threshold selection, which simultaneously considers sensitivity (Recall) and specificity.

3 Structural Model of PE Files

Portable Executable (PE) is the standard format for executable files in Windows operating systems. It describes the structure of .exe, .dll, .sys, and other file types, ensuring correct interaction with the operating system loader.

For static analysis, a formalized description of the PE file is required. Let us denote the set of all valid PE files as F. Validity means compliance with the structure and constraints defined in the Microsoft PE Format specification. Each file $f \in F$ represents an ordered tuple of components:

$$f = (H, S, I, E, R, O) \tag{1}$$

where: H—PE file headers; S—section table; I—import table; E—export table; R—resources; O—overlay.

Based on the analysis of PE file structure and existing research [21], six main feature groups were identified for automating malware detection:

1. Structural anomalies. These include the absence of mandatory headers such as MZ and PE\0\0, mismatch of header and section sizes, overlapping sections, use of RWX flags in sections.

2. Anomalies in import and export tables. Malware often uses specific API calls for code injection, hidden network connections, registry manipulation.

3. Digital signatures. The absence of a digital signature is a serious indicator of potentially malicious software.

4. Use of protection mechanisms. If a PE file disables DEP, ASLR, or CFG, this may indicate exploitation attempts.

5. Anomalies in resources and metadata. PE file resources may contain embedded malware, encrypted data, hidden loaders.

6. Use of specific APIs and functions. These include debugger checks (IsDebuggerPresent), system registry manipulation, and process modification.

We define the PE file model as the set of features (indicators) $I = \{I_1, I_2, \ldots, I_n\}$ – elements of the studied object's structure sufficient for solving the malware detection task by static analysis.

4 Research Methodology

This study addresses the following research questions:

RQ1: What is the optimal feature subset size that balances classification accuracy and computational efficiency for PE file malware detection?

RQ2: Can a cascade classification architecture reduce analysis time without sacrificing detection quality compared to single-level models?

RQ3: What is the critical threshold of malicious traffic share at which cascade classification remains more efficient than direct classification?

The research methodology consists of the following interconnected stages:

Stage 1. A representative sample of 34,026 PE format executable files was collected for analysis. Malicious files were provided by the VirusShare resource [20]. The archive VirusShare_00486.rar dated February 14, 2024, containing 65,536 malware samples was used, of which PE format files comprised 17,992 samples. Legitimate files (16,034) were collected from different Windows versions (7, 8, 10).

Stage 2. Selection of optimal feature sets for the cascade classifier. Based on the preliminary feature importance analysis using the Extra Trees method, two subsets were identified: 10 features for the initial cascade level and 40 features for refining classification.

Stage 3. Comparative analysis of object processing time when using models trained on different feature volumes.

Stage 4. Classification threshold tuning. At the first cascade level, the threshold was selected empirically considering minimization of Type II error (FNR). At the second level, the threshold was determined by Youden's index.

Stage 5. Comprehensive verification of the developed architecture by quality metrics (Precision, Recall, F1) and time characteristics.

5 Results

5.1 Feature Space Optimization

In the experiment, a vector of 333 binary indicators was formed for each PE file. For evaluating the effectiveness of binary representation, a set of machine learning algorithms was used: ensemble methods (Extra Trees, Random Forest, Gradient Boosting, AdaBoost), linear approaches (Logistic Regression, Ridge Classifier, LDA), and k-nearest neighbors method.

Table 1 presents summary accuracy indicators of models using binary feature space.

Table 1. Classification metrics for different algorithms (binary encoding)

Model	Accuracy	Recall	Precision	F1-score
Extra Trees Classifier	**0.965**	0.948	**0.985**	**0.966**
Random Forest Classifier	0.965	0.948	0.985	0.966
Decision Tree Classifier	0.959	**0.949**	0.972	0.960
K Neighbors Classifier	0.954	0.938	0.974	0.956
Gradient Boosting Classifier	0.915	0.896	0.941	0.918

Considering the limitations of binary representation, importance levels for indicators were formed. Each indicator I_k is assigned an importance level w_k—an integer from 1 (minimum) to 9 (maximum).

Bayesian Calibration of Feature Importance Levels. To provide a scientifically grounded methodology for assigning severity levels, a probabilistic interpretation based on Bayes' theorem was introduced. For each binary indicator f_i, the severity is derived from the posterior probability of maliciousness:

$$P(M \mid f_i) = \frac{P(f_i \mid M) \cdot P(M)}{P(f_i \mid M) \cdot P(M) + P(f_i \mid L) \cdot P(L)} \tag{2}$$

where M denotes the malicious class, L the legitimate class, and the prior probabilities are: $P(M) = 17992/34026 \approx 0.529$, $P(L) \approx 0.471$. The conditional probabilities are estimated with Laplace smoothing:

$$\hat{P}(f_i \mid M) = \frac{c_{M,i} + 1}{N_M + 2}, \quad \hat{P}(f_i \mid L) = \frac{c_{L,i} + 1}{N_L + 2} \tag{3}$$

where $c_{M,i}$ and $c_{L,i}$ are the counts of files containing feature f_i in each class. The posterior probability is mapped to the 1–9 severity scale:

$$S_i = \text{round}(1 + 8 \cdot P(M \mid f_i)) \tag{4}$$

For example, "imports WriteProcessMemory API" with $c_M = 4500$, $c_L = 300$ yields $P(M \mid f) \approx 0.937$, resulting in $S = 9$ (critical). Verification on the top-40 features showed high correlation (Pearson $r > 0.9$) between Bayesian severity and model-derived importance scores.

After converting data to this format, classification quality improved across all main metrics: Accuracy increased by an average of 0.02, reaching 0.984 in the best models.

For dimensionality reduction, Principal Component Analysis (PCA) and Isolation Forest methods were used. The results presented in Table 2 demonstrate that 40 most informative features achieve maximum accuracy.

Table 2. Classification quality depending on feature space dimensionality

Features	Accuracy	Precision	Recall	F1-score	FPR	FNR
10	0.914	0.920	0.917	0.918	0.088	0.082
20	0.962	0.965	0.962	0.964	0.038	0.037
30	0.962	0.976	0.979	0.977	0.026	0.020
40	**0.981**	**0.982**	**0.982**	**0.982**	**0.019**	**0.017**
50	0.981	0.982	0.981	0.982	0.019	0.018

5.2 Analysis of Computational Costs

Experimental measurements were performed on five models trained on feature subsets of 10, 20, 30, 40, and 333 features. Table 3 demonstrates a clear dependence of processing time on feature space dimensionality.

Table 3. Average analysis time dependence

Features	Legitimate files (s)	Malicious files (s)	Average (s)
10 features	1.82	2.29	2.06
20 features	3.15	4.81	3.98
40 features	3.38	5.34	4.36
333 features	43.24	121.42	82.33

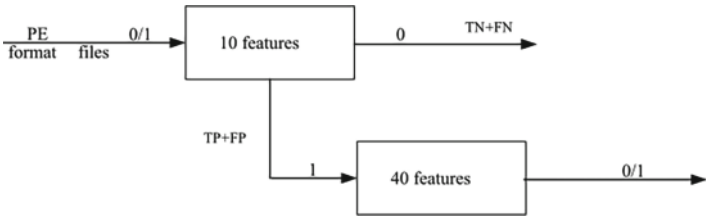


Fig. 1. Cascade classifier structure

5.3 Cascade Classifier Threshold Configuration

The cascade classifier structure is shown in Fig. 1.

For selecting the optimal threshold at the second level (40 features), Youden's J-statistic was used:

$$J = TPR + TNR - 1 \tag{5}$$

where TPR is the true positive rate (Recall), and TNR is the true negative rate $(1 - FPR)$.

The optimal value for the 40-feature model was determined to be 0.54. At this value: Precision = 0.988, Recall = 0.986, FPR = 0.012, FNR = 0.013, with maximum $J = 0.973$.

For the first cascade level (10 features), the operating point is intentionally shifted toward high Recall, accepting a relatively high false positive rate (FPR $\approx$ 40%), since the primary goal at this stage is to avoid missing any malware while filtering out approximately 60% of obviously legitimate files. The optimal threshold value was determined to be 0.16.

Table 4 presents classification model results.

Table 4. Classification model results

Features	Model	Threshold	Precision	Recall	F1-score	α	β
10	DecisionTree	0.16	0.733	0.990	0.842	0.407	0.009
40	RandomForest	0.54	0.988	0.986	0.987	0.012	0.013

5.4 Time Efficiency Calculation and Cascade Applicability Boundary

The cascade operates as follows: if the first-level classifier predicts "Legitimate" (probability below threshold), analysis stops immediately; if it predicts "Malicious" (probability above threshold), the file is forwarded to the second level

for refined classification. The expected average analysis time per file can be expressed as:

$$T_{cascade} = T_{avg}^{10} + T_{avg}^{40} \cdot (P_M \cdot TPR_{10} + P_L \cdot FPR_{10}) \quad (6)$$

where P_M is the share of malicious files in the input stream, $P_L = 1 - P_M$ is the share of legitimate files.

Solving the equation $T_{cascade} = T_{avg}^{40}$, we obtain $P_M = 0.206$, corresponding to a critical share of malicious objects in the input stream of 20.6%. Thus, at $P_M < 20.6\%$, the cascade architecture provides time savings compared to the 40-feature model.

Table 5 presents efficiency analysis results.

Table 5. Cascade architecture efficiency results

Malware share, P_M	Legitimate share, P_L	$T_{cascade}$ (s)	Time savings (%)
0.00	1.00	3.834	12.05
0.01	0.99	3.860	11.47
0.10	0.90	4.089	6.21
0.20	0.80	4.344	0.37
0.50	0.50	5.108	−17.15

The results of this analysis are illustrated in Fig. 2.

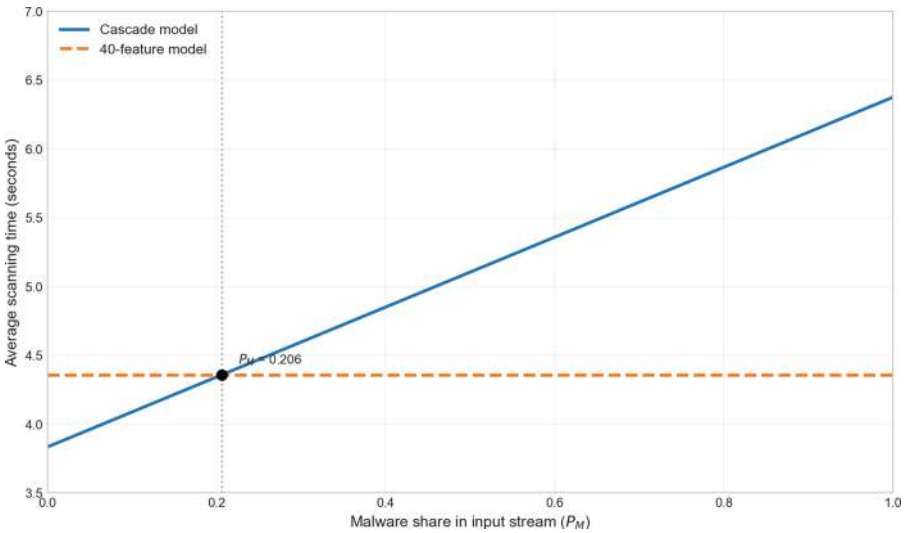


Fig. 2. Scanning time dependence on malware share

6 Discussion

The developed two-level cascade architecture for static classification of PE files demonstrated consistent performance in both quality and computational metrics. At the first cascade level, a decision tree trained on ten most informative features is applied. At the empirically selected threshold of 0.16, the model provides high recall of 0.990 with moderate precision of 0.733.

The second level is implemented by an ensemble performing refining classification with threshold 0.54. This combination provides a final F1-score of 0.987 with average analysis time per PE file of about 4 s.

Compared to single-level models, the cascade approach provides equivalent recognition quality while reducing average processing time by 5–12% in the range of malicious object shares up to 20%.

Tree-based models (Decision Tree and Random Forest) were chosen over deep learning approaches due to their faster inference time on CPU and greater interpretability, which is crucial for integration into endpoint protection systems where computational resources are limited.

7 Conclusion

This study developed a two-level cascade PE file classifier and provided answers to the posed research questions.

Regarding **RQ1**, experiments demonstrated that 40 features represent the optimal subset size, achieving maximum classification quality (F1-score 0.982) with a 19-fold reduction in analysis time compared to the full 333-feature set. Further increase in features does not improve accuracy but increases computational costs.

For **RQ2**, the cascade architecture combining a Decision Tree (10 features, threshold 0.16) at the first level and Random Forest (40 features, threshold 0.54) at the second level achieves equivalent recognition quality (F1-score 0.987) while reducing average processing time by 5–12% compared to direct 40-feature classification.

Answering **RQ3**, mathematical analysis established that the critical malicious traffic share is $P_M = 20.6\%$. Below this threshold, the cascade provides time savings; above it, direct classification becomes more efficient.

The classifier design is based on rational differentiation of feature space, where the primary decision is made by a model on 10 static features, while in-depth verification is performed by a model on 40 features. This approach is complemented by a formalized threshold tuning procedure based on Youden's index, ensuring the required ratio of Type I and Type II errors.

Prospects for further work include automatic threshold adjustment for changing class ratios and using dynamic features, which will further improve detection reliability while maintaining achieved performance.

Conflict of Interest The authors declare no conflict of interest.

Data Availability The malware dataset used in this study is available from VirusShare.com [20] upon registration. The legitimate PE files were collected from official Windows installations.

References

1. Schultz, M.G., Eskin, E., Zadok, E., Stolfo, S.J.: Data mining methods for detection of new malicious executables. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 38–49 (2001). https://doi.org/10.1109/SECPRI.2001.924286
2. Yuk, C.K., Seo, C.J.: Static analysis and machine learning-based malware detection system using PE header feature values. Int. J. Innov. Res. Sci. Stud. **5**(4), 281–288 (2022). https://doi.org/10.53894/ijirss.v5i4.690
3. Gao, J., Wang, J., Fang, Z., Zhang, Y., Wu, D., Ge, W.: A survey of strategy-driven evasion methods for PE malware: transformation, concealment, and attack. Comput. Secur. **137**, 103595 (2024). https://doi.org/10.1016/j.cose.2023.103595
4. García, D.E., DeCastro-García, N.: Optimal feature configuration for dynamic malware detection. Comput. Secur. **105**, 102250 (2021). https://doi.org/10.1016/j.cose.2021.102250
5. Yousuf, M.I., Anwer, I., Riasat, A., Zia, K.T., Kim, S.: Windows malware detection based on static analysis with multiple features. PeerJ Comput. Sci. **9**, e1319 (2023). https://doi.org/10.7717/peerj-cs.1319
6. Chen, Z., Zhang, X., Kim, S.: A learning-based static malware detection system with integrated feature. Intell. Autom. Soft Comput. **27**, 891–908 (2021). https://doi.org/10.32604/iasc.2021.016933
7. Baker del Aguila, R., Contreras-Pérez, C.D., Silva-Trujillo, A.G., Cuevas-Tello, J.C., Nunez-Varela, J.: Static malware analysis using low-parameter machine learning models. Computers **13**(3), 59 (2024). https://doi.org/10.3390/computers13030059
8. Saleh, M.A.: Malware detection approaches based on operation codes (OpCodes) of executable programs: a review. Indones. J. Electr. Eng. Inform. **11**(2), 570–585 (2023). https://doi.org/10.52549/ijeei.v11i2.4454
9. Samantray, O.P., Tripathy, S.N.: An opcode-based malware detection model using supervised learning algorithms. Int. J. Inf. Secur. Priv. **15**(4), 18–30 (2021). https://doi.org/10.4018/IJISP.2021100102
10. Yeboah, P.N., Amuquandoh, S.K., Musah, H.B.B.: Malware detection using ensemble n-gram opcode sequences. Int. J. Interact. Mob. Technol. **15**(24), 19–31 (2021). https://doi.org/10.3991/ijim.v15i24.25401
11. Abusitta, A., Li, M.Q., Fung, B.C.M.: Malware classification and composition analysis: a survey of recent developments. J. Inf. Secur. Appl. **59**, 102828 (2021). https://doi.org/10.1016/j.jisa.2021.102828
12. Wu, X., Song, Y., Hou, X., Ma, Z., Chen, C.: Deep learning model with sequential features for malware classification. Appl. Sci. **12**(19), 9994 (2022). https://doi.org/10.3390/app12199994
13. Zhu, X., Huang, J., Wang, B., Qi, C.: Malware homology determination using visualized images and feature fusion. PeerJ Comput. Sci. **7**, e494 (2021). https://doi.org/10.7717/peerj-cs.494

14. Damaševičius, R., Venčkauskas, A., Toldinas, J., Grigaliūnas, Š: Ensemble-based classification using neural networks and machine learning models for Windows PE malware detection. Electronics **10**(4), 485 (2021). https://doi.org/10.3390/electronics10040485
15. Kim, S., Yeom, S., Oh, H., Shin, D., Shin, D.: Automatic malicious code classification system through static analysis using machine learning. Symmetry **13**(1), 35 (2021). https://doi.org/10.3390/sym13010035
16. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE CVPR 2001, vol. 1, pp. I-511–I-518. IEEE (2001). https://doi.org/10.1109/CVPR.2001.990517
17. Mahato, A., Majumdar, R., Ghosh, S.K.: Feature-driven malware detection using cascade machine learning models. SN Comput. Sci. **6**(7), 794 (2025). https://doi.org/10.1007/s42979-025-04342-1
18. Canbek, G., Temizel, T.T., Sagiroglu, S.: PToPI: a comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. SN Comput. Sci. **4**, 13 (2022). https://doi.org/10.1007/s42979-022-01409-1
19. Schubert Kabban, C.M., Graham, S.R.: Malware classification through abstract syntax trees and L-moments. Comput. Secur. **133**, 104082 (2025). https://doi.org/10.1016/j.cose.2024.104082
20. VirusShare.com: A collection of malware samples for research purposes. https://virusshare.com/ (Accessed: 30.11.2025)
21. Cohen, A., Nissim, N., Rokach, L., Elovici, Y.: SFEM: structural feature extraction methodology for the detection of malicious office documents using machine learning methods. Expert Syst. Appl. **64**, 324–338 (2016). https://doi.org/10.1016/j.eswa.2016.07.010
22. Muralidharan, T., Cohen, A., Gerson, N., Alazab, M.: File packing from the malware perspective: techniques, analysis approaches, and directions for enhancements. ACM Comput. Surv. **55**(5), 108 (2023). https://doi.org/10.1145/3530810
23. Saxe, J., Berlin, K.: Deep neural network-based malware detection using two-dimensional binary program features. In: 10th International Conference on Malicious and Unwanted Software (MALWARE), pp. 11–20. IEEE (2015). https://doi.org/10.1109/MALWARE.2015.7413680
24. Ucci, D., Aniello, L., Baldoni, R.: Survey of machine learning techniques for malware analysis. Comput. Secur. **81**, 123–147 (2019). https://doi.org/10.1016/j.cose.2018.11.001
25. Aslan, O., Samet, R.: A comprehensive review on malware detection approaches. IEEE Access **8**, 6249–6271 (2020). https://doi.org/10.1109/ACCESS.2019.2963724

Developing the Readiness of Future Teachers to Use Artificial Intelligence in Their Professional Activities

Tatyana Kulikova[1](✉), Natalya Poddubnaya[1], Olga Fedina[1], and Prakasha G S[2]

[1] North-Caucasian Federal University, Pushkin Street, 1, Stavropol, Russia
t_a_kulikova@mail.ru
[2] School of Education, Christ University, Bengaluru, India

Abstract. The article analyzes the work of researchers on the problems of applying AI capabilities in education. The necessity of training future subject teachers to use AI in their professional activities is substantiated. A model for the formation of a future teacher's readiness to use artificial intelligence tools in accordance with their didactic capabilities is proposed, which includes the following modules: familiarization with the didactic capabilities of AI; study and selection of AI tools for the implementation of didactic goals; development of digital content based on AI; methodological techniques for using the developed digital materials in the educational process. The implementation of the developed model is aimed at teaching students to create an integrated educational ecosystem in which digital technologies do not replace, but complement and expand the capabilities of human resources. A variant of the implementation of this learning model is proposed, which has been tested and implemented in the educational process of the bachelor's degree in pedagogical fields on the basis of the North Caucasus Federal University. A survey of students conducted before the start of the experiment revealed the desire of students to study the feasibility and effectiveness of introducing AI into teaching practice, as well as to gain knowledge and skills of using AI in their future professional activities. The experiment made it possible to assess the initial level of knowledge of students, as well as to record changes in their knowledge, skills and abilities that occurred as a result of training. The data obtained showed the effectiveness of the proposed learning model. A survey of students conducted at the end of the experiment allows us to conclude that the purposeful use of the proposed technology ensures the effectiveness of the learning process, helps maintain constant cognitive interest and increase motivation for learning activities. Students note that the competent use of AI in their further teaching activities will allow them to entrust routine work to machines and open up opportunities for greater communication with students and engaging in creative activities.

Keywords: Artificial Intelligence in education · Methodological techniques · Educational environment · AI tools · Didactic capabilities of AI · Digital content development · Generative AI · Model

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 274–288, 2027.
https://doi.org/10.1007/978-3-032-27834-0_26

1 Introduction

In the context of the transformation of the modern education system, the use of artificial intelligence (AI) technologies to improve teaching methods and forms and, in general, to increase the effectiveness of a teacher's professional activities is of increasing interest.

The use of artificial intelligence (AI) in teaching and learning is innovative for the field of education. Artificial intelligence refers to the field of computer science that focuses on creating programs and systems that can analyze data, learn from experience, and make decisions that would seem reasonable if made by a human. This includes machine learning, artificial neural networks, natural language processing algorithms, pattern recognition (computer vision), and more [1]. Artificial neural networks (neural networks) are one of the methods used to implement artificial intelligence (AI). In fact, AI mimics human thought processes [2, 3]. In the context of education, AI is a branch of computer science that allows machines to perform tasks traditionally considered the prerogative of the human mind, such as logical thinking, decision-making, and learning [4, 5].

The educational environment, in turn, covers all aspects – physical, social and technological – in which the learning process takes place [6]. This environment has a significant impact on how students acquire knowledge and skills, due to factors such as accessible infrastructure, educational materials, and active participation in the learning community. Zysberg and Schwabsky [7] define the educational environment as a reflection of society, contributing to the education of citizenship and the formation of a sense of community. Tetzlaff and his colleagues [8] emphasize that this environment is dynamic and adaptable. In the age of AI, the interaction between societal values and the flexibility of the environment is especially important, as educational institutions must now implement intelligent systems for a more personalized, accessible, and generally improved educational experience [9].

The role of AI in education can be analyzed from two key theoretical perspectives: constructivism and connectivism. Constructivism views learning as an active process in which students build knowledge through interaction with their environment and collaboration with others [10]. AI fits perfectly into this model by providing adaptive and interactive tools that facilitate experimental and independent (personalized) learning [11]. Similarly, Connectivism emphasizes that knowledge is distributed across networks and that learning occurs by navigating these connections [12]. AI acts as a critical node within these networks, facilitating access to diverse sources of information and promoting collaboration [13, 14].

2 Generative AI Tools and Their Advantages in Education

Over the past decade, scientific research has increasingly pointed to the significant benefits of integrating artificial intelligence (AI) into the educational process. Special attention is paid to the ability of AI to personalize learning, stimulate student interest and activity, and serve as methodological support for educators [15, 16]. AI proponents emphasize that its applications add additional value to learning, opening up innovative opportunities for students and enriching the educational environment as a whole [17].

Thanks to various AI tools, existing pedagogical approaches can be adapted to meet diverse learning needs and improve overall academic performance [18, 19].

In this new, AI-oriented educational paradigm, the roles of educators require rethinking in the light of recent technological advances and continuous societal changes. This transformation poses a number of challenges for teachers: they need to understand their new role in the educational process and harmoniously integrate their professional responsibilities with the concept of AI [20]. At the same time, educators are concerned about the appropriate and effective implementation of AI in teaching practice and curricula in order to maximize the benefits for students.

The Council of Europe report «Artificial Intelligence and Education» systematizes research in the field of artificial intelligence in education (AIEd) in four areas [21].

The use of AI in pedagogical practice. This area covers research aimed at the direct application of AI in the educational process to support students' learning activities, as well as the use of AI systems to optimize the administrative functions of teachers and the introduction of new pedagogical approaches.

Analysis of the learning process using AI. This category of research focuses on the application of AI to study and analyze the dynamics of the educational process in order to identify patterns and trends in student learning, which, in turn, can contribute to the development of pedagogical strategies.

An exploration of the nature and principles of AI. This area of research is aimed at deepening fundamental knowledge about artificial intelligence, its functioning mechanisms, as well as at developing optimal methods for teaching AI and effective strategies for spreading knowledge about it.

Preparing for the integration of AI into society. This area of research is engaged in the development of approaches to preparing people for life and professional activity in an AI-dominated environment. Unlike studying the technical aspects of AI, this field focuses on developing the skills of the population to effectively apply AI in various areas of life and work.

Generative AI, due to its ability to create seemingly original content from existing data, has a transformative impact on the modern world. Feuerriegel et al. describe it as computational methods capable of generating seemingly new, meaningful content from training data [22]. A prime example in the field of natural language processing is one that can process, understand, and generate text that mimics human language based on user queries. Such tools, including ChatGPT and other advanced AI systems, open up new opportunities for the development of training programs and educational courses [23].

Over the past decade, AI proponents have predicted its potential to implement innovative ways to gain knowledge. Moreover, AI can dramatically change the way students are given feedback and evaluated, offering personalized and automated solutions that affect the overall assessment of academic performance.

In addition, the personalized feedback provided by many AI tools helps teachers identify areas that require additional attention and motivate students. According to Spector and Ma [24], AI stimulates critical thinking by offering students a variety of tasks and opportunities for innovative practice. Research [25] also shows that AI can positively influence students' reading and math skills.

Current studies are focusing primarily on in-service teachers in teacher education [26], and only a few studies deal with the question of how teacher training can be improved to prepare pre-service teachers for the future teaching implementation of AI [27]. As a consequence, the TPACK model has been extended to describe the technological and pedagogical content knowledge required for effectively integrating AI in teaching, resulting in the AI-TPACK model [28].

In our opinion, it is necessary to consider the training of future teachers to use AI competently, both from a technological, pedagogical, and methodological perspective. The proposed model will help to solve these issues.

3 Model for the Formation of the Readiness of the Future Teacher to Use Artificial Intelligence Tools in Professional Activities

In our opinion, the model of formation of a future teacher's readiness to use artificial intelligence tools in the learning process should be represented by the following modules (see Fig. 1): familiarization with the didactic capabilities of AI, study and selection of AI tools for the implementation of didactic goals, development of AI-based digital content, methodical receptions for using the developed digital materials in the educational process.

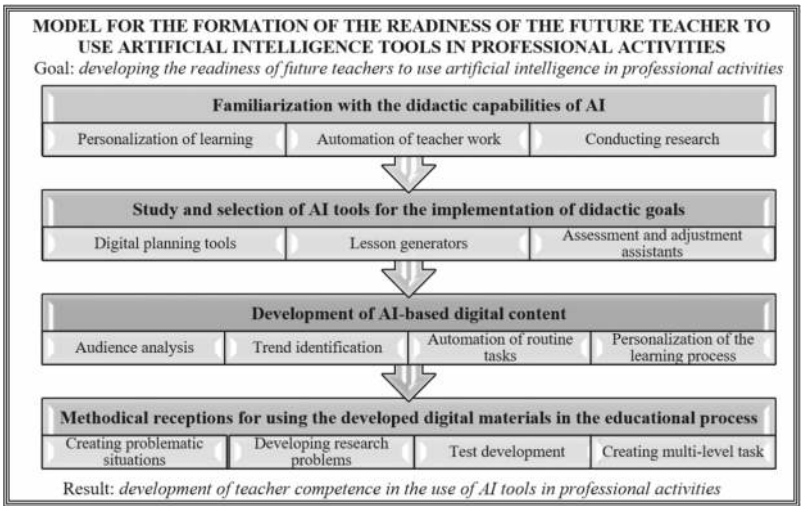


Fig. 1. Model for the formation of the readiness of the future teacher to use artificial intelligence tools in professional activities.

Let's look at the selected modules in more detail.

3.1 Didactic Capabilities of AI

In the first module, students get acquainted with the didactic potential of artificial intelligence, which allows them to personalize learning, create interactive gamified content,

automate the routine work of teachers when checking tests, developing differentiated assignments, planning and implementing research papers, students, etc.

Figure 2 shows the structure of the module, which highlights the didactic capabilities of AI, defines their educational goals and objectives, analyzes and suggests AI tools.

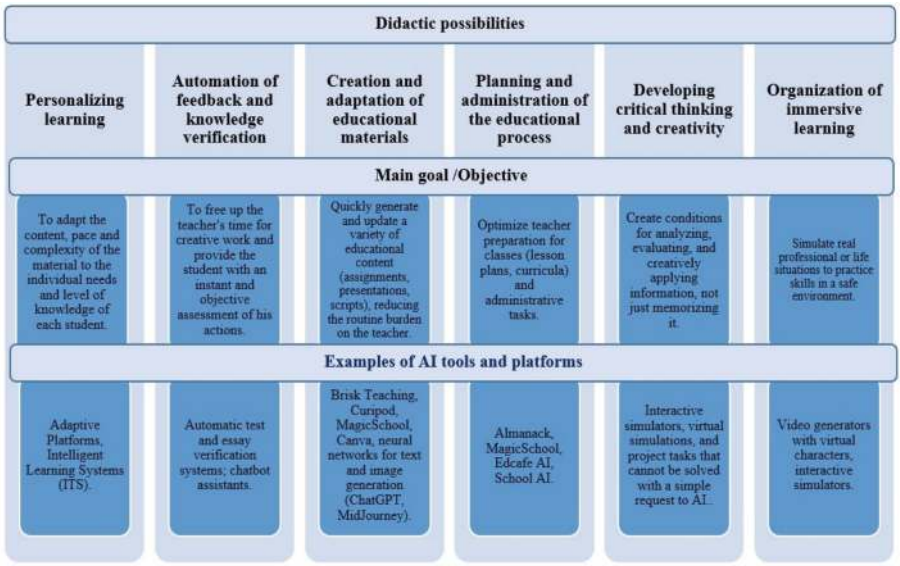


Fig. 2. The structure of the module «Familiarization with the didactic capabilities of AI».

Due to its didactic properties, artificial intelligence opens the way to building a flexible, convenient and effective educational environment. To achieve success in its application, a competent interaction of technological tools and the professional and human aspect of the teacher's activity is required. This is the object of study of the second module.

3.2 Study and Selection of AI Tools for the Implementation of Didactic Goals

In order to achieve a significant increase in the effectiveness of learning using artificial intelligence (AI), the teacher needs to realize its complex potential. AI does not just automate routine tasks, but contributes to the creation of a holistic educational ecosystem where digital technologies harmoniously complement the human factor, empowering teachers and students. When choosing digital tools with AI, a teacher should be able to identify those that are most relevant for each stage of the educational process. Students master educational materials for the study and competent selection of AI tools for the implementation of didactic learning goals in the second module.

The pedagogical process is traditionally divided into three successive stages: preparatory, basic and final. For each of them, there are digital solutions supported by neural networks.

One of them may be, for example, digital planning tools based on AI.

At the preparatory stage, the teacher formulates specific learning objectives, analyzes learning conditions, predicts possible outcomes, and on this basis develops a plan for the further development of the educational process [29].

Neural networks are capable of creating high-quality illustrations for lessons, projects, and presentations. Consider several AI-powered digital tools designed to create unique, high-quality images. Kandinsky is a neural network that allows you to create photorealistic images with higher resolution. You can also use it to fill in missing parts of an image and create paintings in the infinite canvas mode. Dezgo and Lexica are neural networks focused on creating images from text descriptions and converting any image into a high-quality picture.

Each neural network allows you to generate unique content and allows you to refine and modify the details of the image according to the textual request. However, when using tools for creating visual aids, it is important to consider the relevance of the generated images to the learning objectives. Any illustrative materials, including those generated by neural networks, require systematic evaluation and selection. Mistakes in the selection of illustrations can negatively impact the learning experience for students.

Curipod allows you to transform any documents or presentations into interactive lessons, ready for use in learning. The process of creating educational materials is simplified as much as possible. If necessary, you can specify the topic of the lesson and select the required parameters so that Curipod can independently generate the structure and content of the lesson, or upload existing documents or presentations that will form the basis for future lessons.

Another advanced platform that uses the capabilities of artificial intelligence is Almanack. It allows you to intuitively create a structure for various types of classes, enriching them with a variety of educational materials: from slides and videos to worksheets, interactive practical tasks, test materials and game applications.

MagicSchool provides teachers with access to an extensive library of AI-based tools designed to address a variety of learning challenges. These include automated lesson plan generators, test designers, and intelligent assistants. Special mention should be made of the specialized chatbot Raina, developed taking into account the needs of teachers. With its help, the teacher can formulate a request, and the bot will provide valuable ideas and practical recommendations.

Another innovative AI assistant for teachers and students is the Roshi platform, which significantly simplifies the process of designing and developing educational materials and allows you to transform any online content into ready-to-use digital educational resources.

3.3 AI-Based Digital Content Development

As part of the third module, «Development of AI-based digital content» students learn how to create various types of educational materials using the AI tools they have learned. For example, tests, keys and criteria for their assessment, differentiated tasks, automatic verification tools, tasks based on typical student errors, etc. It should be noted that before

creating digital neural network content, it is necessary to download theoretical material (textbook, lecture text, methodological guide, etc.), according to which educational materials will be generated.

Let's look at examples of projects and the results of training materials generated by the neural network.

Creating test tasks in ChatGPT (GPT-5).

Promt: Create 10 test tasks of different types on the topic «Text information processing», grade 7. Include different types of tasks:

Closed test tasks (with one correct answer).

Multiple choice assignments (multiple correct choices).

Compliance tasks (concepts, definitions, examples, terms).

Tasks with omissions (it is necessary to insert the missing words/numbers into the formula or definition).

Open-ended questions (short answer or reasoning).

Practical tasks (solving a topic with calculations or an explanation).

At the end, add the keys with the correct answers for verification.

Form it in an easy-to-print form (numbered list).

The result of the ChatGPT neural network (GPT-5), the generated training materials, is shown in Fig. 3.

Test tasks on the topic: "Processing test information," grade 7

1. *Assignment for establishing compliance*. Map the concept to the definition:

1. Concept	a) Definition
2. Paragraph	b) Title of the document section
3. Title	c) Part of text between two indents
4. Symbol	d) Minimum Text Unit

2. *Task with omissions*. Insert missing words:

A text document can contain not only text, but also _____, _____, diagrams.

Test documents are processed using a computer using a _____ editor.

3. *Practical task*

Suppose a page of text contains 40 lines of 60 characters each. How many characters does a 5-page document contain?

Solution: _____

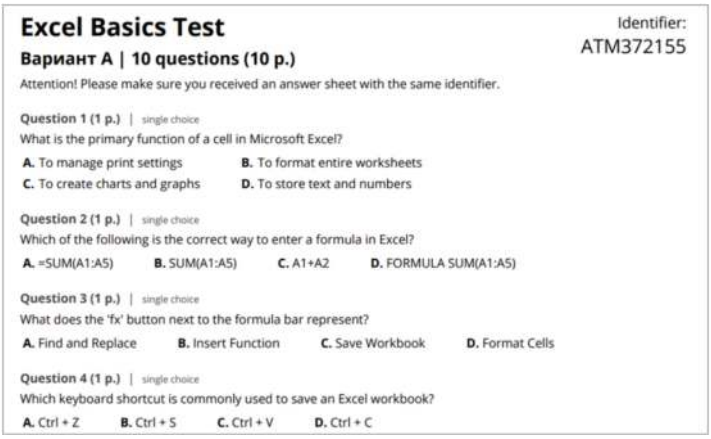
Fig. 3. Test in ChatGPT.

Creation and automatic verification of test assignments in Examica.

The Examica neural network allows teachers to create educational materials to assess students' knowledge and automatically check completed work using a smartphone. This allows you to quickly evaluate the results of mastering the material and reduce the verification time.

To create a neural network test, you must specify the topic and the number of questions. The teacher can correct the generated test. An example of the developed test on the topic: «Basics of working in Excel» is shown in Fig. 4.

You can take the test online by sending students an access link, or offline. The service makes it possible to upload options as a PDF file. Each student is given a number to write on the answer form. An example of the response form is shown in Fig. 5.



Excel Basics Test

Identifier: ATM372155

Вариант А | 10 questions (10 p.)

Attention! Please make sure you received an answer sheet with the same identifier.

Question 1 (1 p.) | single choice

What is the primary function of a cell in Microsoft Excel?

A. To manage print settings **B.** To format entire worksheets

C. To create charts and graphs **D.** To store text and numbers

Question 2 (1 p.) | single choice

Which of the following is the correct way to enter a formula in Excel?

A. =SUM(A1:A5) **B.** SUM(A1:A5) **C.** A1+A2 **D.** FORMULA SUM(A1:A5)

Question 3 (1 p.) | single choice

What does the 'fx' button next to the formula bar represent?

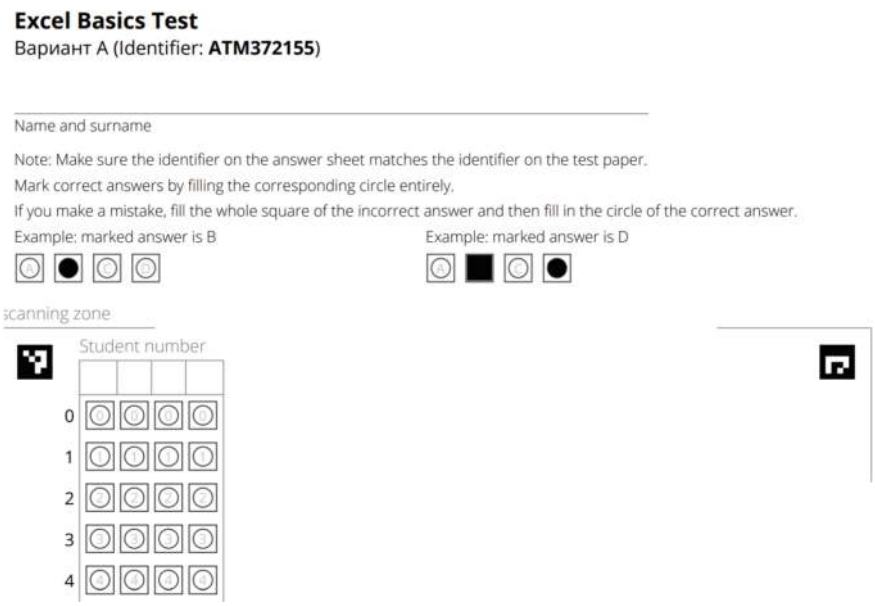
A. Find and Replace **B.** Insert Function **C.** Save Workbook **D.** Format Cells

Question 4 (1 p.) | single choice

Which keyboard shortcut is commonly used to save an Excel workbook?

A. Ctrl + Z **B.** Ctrl + S **C.** Ctrl + V **D.** Ctrl + C

Fig. 4. Test in Examica.



Excel Basics Test

Вариант А (Identifier: **ATM372155**)

Name and surname

Note: Make sure the identifier on the answer sheet matches the identifier on the test paper.

Mark correct answers by filling the corresponding circle entirely.

If you make a mistake, fill the whole square of the incorrect answer and then fill in the circle of the correct answer.

Example: marked answer is B

Example: marked answer is D

scanning zone

Student number

0 1 2 3 4

Fig. 5. Examica response form.

You can check the answers using your smartphone by downloading the app on it. The result of the completed test is shown as a percentage.

Creating a visualization in ChatGPT for a problematic situation.

For example, in a physics lesson, when explaining the new topic «Motion of a body thrown at an angle to the horizon» the teacher can visualize a change in the trajectory of the bodies depending on the launch angles to organize a heuristic conversation.

Promt: Plot the flight paths of bodies thrown at an angle of 15°, 30°, 45° and 60° to the horizontal. Air resistance is not taken into account, the initial velocity is 20 m/s.

The result of the ChatGPT neural network (GPT-5) is shown in Fig. 6.

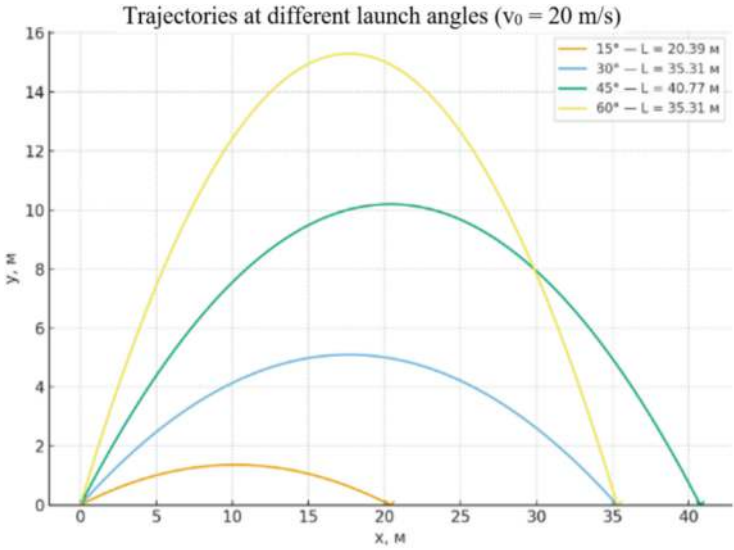


Fig. 6. Image in ChatGPT.

Organizing experimental work to solve a research problem in ChatGPT.

Promt: To develop a plan to study the dependence of the sizes of hexagonal and labyrinthine structures formed in thin layers of ferrofluid on the strength of the magnetic field.

The result of the ChatGPT neural network is shown in Figs. 7 and 8.

The neural network determined not only the experimental plan, offering equipment, with a detailed description of the characteristics, setting the parameters to be measured, but also diagnosed the expected result, warned against possible errors, and suggested ways to verify.

3.4 Methodological Techniques for Using the Developed Digital Materials in the Educational Process

Future teachers need to understand that explaining new material is one of the key moments in a lesson, where teachers must capture students' attention. To do this, they must engage students, sometimes even surprise them with the information. AI can help teachers create problematic situations. For example, when studying a topic related to ballistics, they can pose problematic questions:

1. Which stone will fall farther if thrown at an angle of 20° or 70° to the horizontal? Students typically answer that the first stone will fly farther.

The purpose of the experiment

To investigate how changes in the strength of an external magnetic field affect the dimensions (period, width, height) of structures formed on the surface of a ferrofluid layer with a normal (perpendicular) field orientation.

Equipment and materials

Ferrofluid (e.g., oil-based, with magnetite nanoparticles)
Flat-bottomed container (glass, plastic, or ceramic)
Electromagnet or Helmholtz coil (to create a controlled, uniform magnetic field)
Power supply with adjustable current
Magnetometer/Hall sensor (to measure field strength, H)
Illuminator (to visualize the structure from above)
High-resolution camera/microscope with camera
Image processing software (e.g., ImageJ)
Micrometer ruler or graticule (photo calibration)

Experimental Plan

Step 1. Preparing the Setup

1. Pour a thin layer of ferrofluid (2 – 5 mm) into the container.
2. Place the container between the coils (if using Helmholtz coils) or under the electromagnet.
3. Install a magnetometer to precisely monitor the H voltage.
4. Connect the power source to the electromagnet and adjust the current control.
5. Set up the lighting and camera above.

Fig. 7. Experiment organization process in ChatGPT

Expected results

At **low H**, there is no structure.
At **H ⩾ H_crit**, hexagonal peaks (sometimes "needle-like") form.
With further increases in H, a transition **to labyrinthine structures is possible**.
The size of the structures usually decreases with increasing H (the structures "compress").
A **nonlinear relationship** between H and geometry can be detected.

Important to note:

The temperature must be stable (the viscosity of the ferrofluid is temperature-dependent).
The thickness of the ferrofluid layer also affects the nature of the structures.
Use **distilled water** or oil as a medium (if viscosity variation is required).

Potential sources of error

Magnetic field instability.
Inhomogeneous ferrofluid layer thickness.
Visualization errors (glare, artifacts).
Inability to accurately measure dimensions due to poor contrast.

Fig. 8. Expected results and possible errors of the ChatGPT experiment.

2. At what angle must a stone be thrown for it to fly the farthest?

Unfortunately, not all schools have experimental installations that demonstrate flat and inclined body trajectories. Using AI, you can create a graphical representation of the experiment result and show it to students (see Fig. 6).

Research problem solving is an important part of the physics curriculum. A future teacher should be able to independently formulate research problems and solve them

with students after gathering the necessary teaching materials. Developing research problems is painstaking and complex work that requires considerable time. Using AI will help teachers not only select problems but also solve them relatively quickly. When presenting research to students, teachers must be confident of success. To achieve this, they must first solve the problem themselves and ensure that the result can be achieved in a classroom setting. Otherwise, research failures can lead to students losing interest in research.

Scientific research takes years. Using AI, a teacher can create simple investigations for schoolchildren from already solved scientific problems.

Here's an example of a physics research problem, based on a scientific study that AI helped solve: «If an external magnetic field is directed perpendicular to a ferrofluid layer, then under certain conditions, labyrinthine and hexagonal structures form within it. Investigate how the magnitude of the magnetic field influences the size of the resulting structures».

The problem involves ferrofluid, a material unusual for most schoolchildren. Interest in the research can be enhanced by using AI to obtain information about this material and its practical applications in medicine and engineering.

The AI not only determined the experimental plan, suggesting equipment with a detailed description of its characteristics, and setting the parameters to be measured, but also diagnosed the expected result, warned against possible errors, and suggested verification methods.

However, AI doesn't always offer the correct solution. Therefore, a future teacher needs to learn to critically evaluate the solution offered by AI, correct and formulate prompts.

4 Results

The proposed implementation of a technology for developing future teachers' readiness to use artificial intelligence tools in their professional work was implemented into the educational process for undergraduate students majoring in pedagogy at North Caucasus Federal University as part of the course «Innovative Pedagogical Technologies» Students were surveyed before and after completing the course.

The questions and survey results are presented in Figs. 9 and 10.

The survey revealed, on the one hand, increased student interest in using AI in their future professional activities, but also insufficient knowledge of the educational potential of AI and the ability to apply its tools in light of its teaching capabilities.

The initial and final assessments allowed us to assess students' initial knowledge and document changes in their knowledge, skills, and abilities as a result of the training. The data obtained demonstrated the effectiveness of the proposed learning model, as confirmed by the results of students' acquired knowledge and developed skills (see Fig. 11).

An analysis of the results of the student survey and the level of their acquisition of knowledge, skills and abilities showed that the targeted use of the proposed technology ensures the effectiveness of the learning process, helps maintain constant cognitive interest and increases motivation for educational activities.

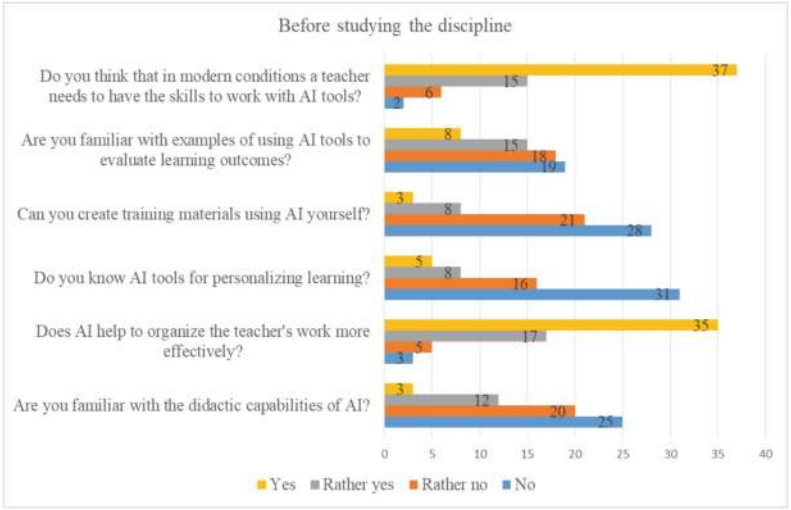


Fig. 9. Results of the entrance survey of students.

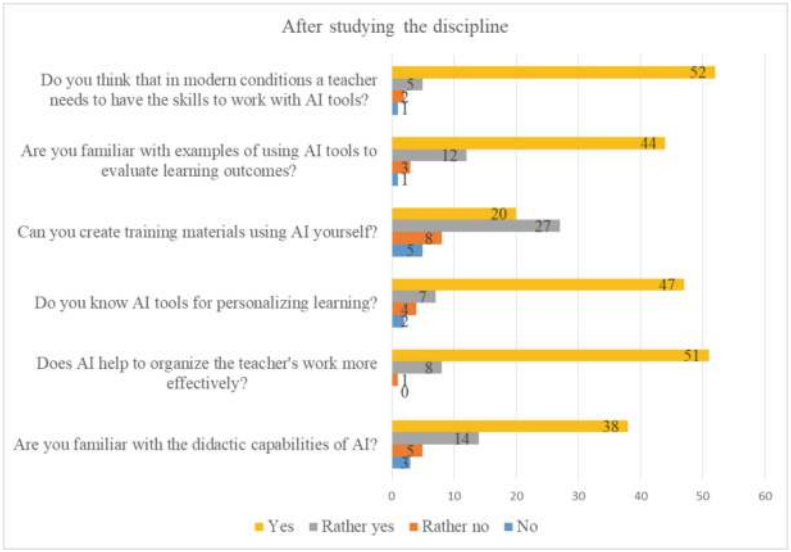


Fig. 10. Results of the final student survey.

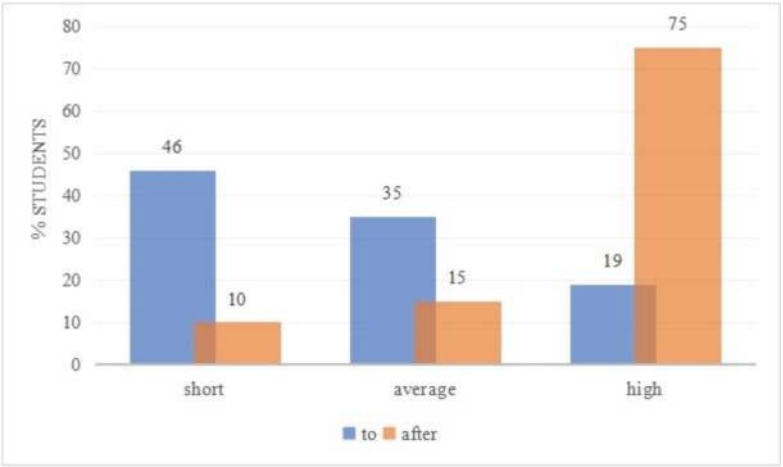


Fig. 11. The level of effectiveness of the learning process

5 Conclusions

In conclusion, artificial intelligence is becoming an integral part of the educational process and an effective teaching assistant. Its use facilitates access to a variety of information resources and promotes collaborative learning, stimulating critical thinking by offering students a variety of challenges and opportunities for innovative practice. For educators, it simplifies the design and development of educational materials and enables the transformation of various content into ready-to-use digital educational resources, thereby optimizing the learning process, increasing teaching effectiveness, and improving learning outcomes.

References

1. Ertel, W.: Introduction to Artificial Intelligence. In: Undergraduate Topics in Computer Science, 3rd edn. Springer, Wiesbaden, Germany (2025). ISBN 978-3-658-43101-3. https://doi.org/10.1007/978-3-658-43102-0
2. Jaboob, A., Durrah, O., Chakir, A.: Artificial intelligence: an overview. In: Engineering Applications of Artificial Intelligence; Synthesis Lectures on Engineering, Science, and Technology, pp. 3–22. Springer, Cham, Switzerland (2024). https://doi.org/10.1007/978-3-031-50300-9_1
3. Markauskaite, L., et al.: Rethinking the entwinement between artificial intelligence and human learning: what capabilities do learners need for a world with AI? Comput. Educ. Artif. Intell. **3**, 100056 (2022). https://doi.org/10.1016/j.caeai.2022.100056
4. Yan, L., Greiff, S., Teuber, Z., Gašević, D.: Promises and challenges of generative artificial intelligence for human learning. Nat. Hum. Behav. **8**, 1839–1850 (2024). https://doi.org/10.1038/s41562-024-02004-5
5. Abulibdeh, A., Zaidan, E., Abulibdeh, R.: Navigating the confluence of artificial intelligence and education for sustainable development in the era of Industry 4.0: challenges, opportunities, and ethical dimensions. J. Clean. Prod. **437**, 140527 (2024). https://doi.org/10.1016/j.jclepro.2023.140527

6. González-Pérez, L.I., Ramírez-Montoya, M.S.: Components of Education 4.0 in 21st century skills frameworks: systematic review. Sustainability **14**, 1493 (2022). https://doi.org/10.3390/su14031493
7. Zysberg, L., Schwabsky, N.: School climate, academic self-efficacy and student achievement. Educ. Psychol. **41**, 467–482 (2021). https://doi.org/10.1080/01443410.2020.1813690
8. Tetzlaff, L., Schmiedek, F., Brod, G.: Developing personalized education: a dynamic framework. Educ. Psychol. **33**, 863–882 (2021). https://doi.org/10.1007/s10648-020-09570-w
9. Irish, A.L., Gazica, M.W., Becerra, V.: A qualitative descriptive analysis on generative artificial intelligence: bridging the gap in pedagogy to prepare students for the workplace. Discov. Educ. **4**, 48 (2025). https://doi.org/10.1007/s44217-025-00435-4
10. Acemoglu, D., Restrepo, P.: Artificial intelligence, automation, and work. In: The Economics of Artificial Intelligence: An Agenda, pp. 197–236. University of Chicago Press, Chicago, IL, USA (2018). https://doi.org/10.1080/00131857.2020.1728732
11. Strielkowski, W., Grebennikova, V., Lisovskiy, A., Rakhimova, G., Vasileva, T.: AI-driven adaptive learning for sustainable educational transformation. Sustain. Dev. **33**, 1921–1947 (2024). https://doi.org/10.1002/sd.3221
12. Goldie, J.G.S.: Connectivism: a knowledge learning theory for the digital age? Med. Teach. **38**, 1064–1069 (2016). https://doi.org/10.3109/0142159X.2016.1173661
13. Gibson, D., Kovanovic, V., Ifenthaler, D., Dexter, S., Feng, S.: Learning theories for artificial intelligence promoting learning processes. Br. J. Educ. Technol. **54**, 1125–1146 (2023). https://doi.org/10.1111/bjet.13341
14. Sperling, K., Stenberg, C.-J., McGrath, C., Åkerfeldt, A., Heintz, F., Stenliden, L.: In search of artificial intelligence (AI) literacy in teacher education: a scoping review. Comput. Educ. Open **6**, 100169 (2024). https://doi.org/10.1016/j.caeo.2024.100169
15. Akgun, S., Greenhow, C.: Artificial intelligence in education: addressing ethical challenges in K-12 settings. AI Ethics **2**(3), 431–440 (2022). https://doi.org/10.1007/s43681-021-00096-7
16. Chiu, T.K.F.: A holistic approach to the design of artificial intelligence (AI) education for K-12 schools. TechTrends **65**, 796–807 (2021). https://doi.org/10.1007/s11528-021-00637-1
17. Seldon, A., Abidoye, O.: The Fourth Education Revolution: Will Artificial Intelligence Liberate or Infantilise Humanity, p. 370. University of Buckingham, Buckingham (2018). ISBN: 978-1908684950. https://doi.org/10.1007/s10734-020-00506-5
18. Holmes, W., Porayska-Pomsta, K., Holstein, K., et al.: Ethics of AI in education: towards a community-wide framework. Int. J. Artif. Intell. Educ. **32**, 504–526 (2022). https://doi.org/10.1007/s40593-021-00239-1
19. Zawacki-Richter, O., Marín, V.I., Bond, M., et al.: Systematic review of research on artificial intelligence applications in higher education—where are the educators? Int. J. Educ. Technol. High. Educ. **16**, 39 (2019). https://doi.org/10.1186/s41239-019-0171-0
20. Fenwick, T.: Pondering purposes, propelling forwards. Stud. Contin. Educ. **40**(3), 367–380 (2018). https://doi.org/10.1080/0158037X.2018.1425205
21. Council of Europe: Artificial intelligence and education. A critical view through the lens of human rights, democracy and the rule of law. https://bluesyemre.files.wordpress.com/2022/11/artificialintelligence-and-education.pdf (2022)
22. Feuerriegel, S., Hartmann, J., Janiesch, C., et al.: Generative AI. Bus. Inf. Syst. Eng. **66**, 111–126 (2024). https://doi.org/10.1007/s12599-023-00834-7
23. Mogavi, R.H., et al.: ChatGPT in education: a blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. Comput. Hum. Behav. Artif. Hum. **2**(1), 100027 (2024). https://doi.org/10.1016/j.chbah.2023.100027
24. Spector, J.M., Ma, S.: Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence. Smart Learn. Environ. **6**, 8 (2019). https://doi.org/10.1186/s40561-019-0088-z

25. Kazakoff, E.R., Macaruso, P., Hook, P.: Efficacy of a blended learning approach to elementary school reading instruction for students who are English Learners. Educ. Tech. Res. Dev. **66**, 429–449 (2018). https://doi.org/10.1007/s11423-017-9565-7
26. Zhang, C., Schießl, J., Plößl, L., et al.: Acceptance of artificial intelligence among pre-service teachers: a multigroup analysis. Int. J. Educ. Technol. High. Educ. **20**, 49 (2023). https://doi.org/10.1186/s41239-023-00420-7
27. Ayanwale, M.A., Sanusi, I.T., Adelana, O.P., Aruleba, K.D., Oyelere, S.S.: Teachers' readiness and intention to teach artificial intelligence in schools. Comput. Educ. Artif. Intell. **3**, 100099 (2022). https://doi.org/10.1016/j.caeai.2022.100099
28. Runge, I., Hebibi, F., Lazarides R.: Acceptance of pre-service teachers towards artificial intelligence (AI): the role of AI-related teacher training courses and AI-TPACK within the technology acceptance model. Educ. Sci. **15**(2), 67 (2025). https://doi.org/10.3390/educsci15020167
29. Kulikova, T.A., Poddubnaya, N.A.: Training of the Future Teacher in Applying Digital Tools for the Organization of Project Activities. Selected Papers of the V All-Russian scientific and practical conference with international participation «Distance Learning Technologies» (DLT 2020), Yalta, Crimea, September 22–25, pp. 391–397. http://ceur-ws.org/Vol-2914/paper37.pdf (2020)

Methods of Identifying Fake News in the Information Space

Vitalii Lapin[1(✉)], Alina Bagautdinova[1], Sergey Fortunatov[2], Alisa Kushschenko[1], and Mbuyu Sumbwanyambe[3]

[1] North Caucasus Federal University, Stavropol, Russia
vitlx@yandex.ru, bagaauttdinova@mail.ru
[2] Moscow University of Finance and Law, Moscow, Russia
[3] University of South Africa, Pretoria, South Africa

Abstract. The problem of fake news is more pressing than ever, posing real threats to society. This article is an attempt to understand how to effectively identify false information. We approach this task from multiple angles, examining high-tech methods like machine learning, traditional expert analysis, and crowdsourced verification. We delve particularly deeply into the technical aspects of the issue—how algorithms that learn to distinguish fact from fiction work, and how linguistic text analysis works. The article focuses on our own development, a web service for automatically detecting fake news. We explain its internal structure, from its architecture and the TAIGA database used to the specific steps of text processing. We explain in detail how ordinary words are transformed into digital data understandable by a computer, and how the model decides whether a text is reliable. The practical value of this work lies in the fact that the proposed approach makes it possible to create a working tool for combating disinformation, combining computer speed with an understanding of linguistic patterns.

Keywords: Machine learning · Information verification · Text classification · Linguistic analysis · Media analytics · Information security · Artificial intelligence · Fake news detection

1 Introduction

Developing robust methods for detecting fake content has become one of the key challenges of the digital age [1]. Traditional approaches to the sudden reliability of information, based on metadata or discussions with trusted sources, often prove ineffective against synthetic media because they fail to take into account the specifics of how such materials are generated. Unlike simple editing, modern deep neural networks create holistic archives that preserve the statistical patterns of the original recordings. This opens up opportunities for the application of machine learning algorithms, demonstrating high accuracy when working with multimodal data. For example, convolutional networks successfully detect artifacts in the frequency domain, and transformers analyze temporal inconsistencies in video [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 289–298, 2027.
https://doi.org/10.1007/978-3-032-27834-0_27

Practical implementation of systems on web platforms is challenging, including the need to process heterogeneous content under strict time constraints. Bandwidth must be able to process thousands of pieces of content per second while maintaining a low false-positive rate [3].

The research will include a comparative analysis of scientific publications over the past three years, experimentally testing hypotheses about combining optical flow and spectral analysis, and assessing reliability thresholds for various neural network architectures. The obtained data will improve external detectors for creating an ensemble of models adaptable to new generative algorithms.

Successful implementation of the project could contribute to ensuring sovereignty and minimizing socio-political risks associated with the widespread distribution of deepfakes [4]. Of particular interest is the development of a universal framework capable of integrating with modern network adapters without increasing latency when processing user content. A promising direction is the creation of hybrid systems combining predictive analytics and active learning methods for the proactive discovery of new types of synthetic environments [5–11].

2 Methodology

The architecture of the developed platform is implemented as four interconnected modules. The educational data collection unit accumulates information from heterogeneous sources: it records the results of interactive tasks, the timing of linguistic problem solving, audio tracks of spoken speech with 44.1 kHz sampling, and interface interaction metrics such as response latency and the frequency of corrections to input responses. When integrated with compatible devices, the system additionally records user biometric parameters [12].

The main components of CNN are:

- Convolutional Layer: Applies filters to input data to extract features.
- Pooling Layer. Reduces the dimensionality of the data while preserving important features. The most common method is max pooling.
- Fully Connected Layer: Used for classification or regression based on extracted features.

CNNs are widely used in computer vision tasks such as image classification, object detection and segmentation.

Figure 1 shows that high resolution (resolution) corresponds to the average level of semantic value (semantic value), while low resolution corresponds to the extreme values - low or high - of semantic value.

Recurrent Neural Networks (RNNs) are designed to process sequences of data, such as time series or text. Unlike CNNs, RNNs have internal memory that allows them to take previous states into account when processing the current element of the sequence [13].

Key features of RNN:

- Hidden State: Stores information about previous elements of the sequence.
- Cyclic connections: Allow information to be transferred from one step to the next.

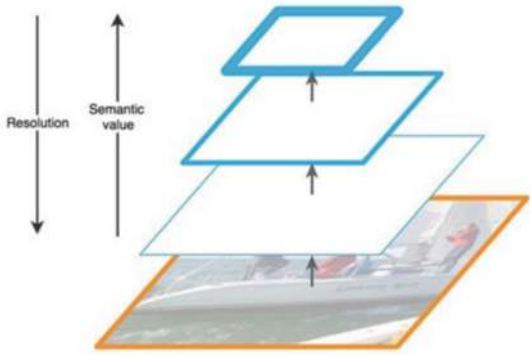


Fig. 1. An example of a CNN architecture for detecting defects.

However, classical RNNs suffer from the vanishing gradient problem, which makes it difficult to learn on long sequences. To address this problem, modifications such as Long Short-Term Memory (LSTM) and Gated Recurent Units (GRU) have been developed.

Figure 2 illustrates the LSTM (Long Short-Term Memory) architecture used for temporal dependency analysis. It schematically depicts an LSTM cell with its key components: the input data structure, hidden state, and information flow control mechanisms (gates), which are represented by arrows illustrating the connections and direction of data flow between network elements.

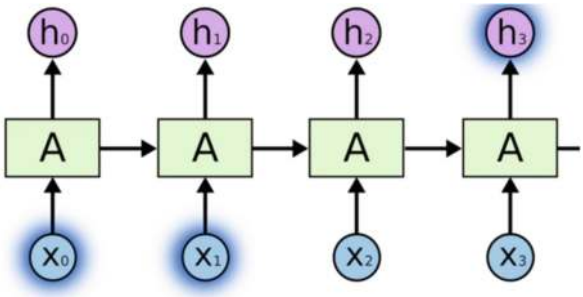


Fig. 2. An example of an LSTM architecture for analyzing time dependencies.

Transformers are an architecture that has revolutionized the field of Natural Language Processing (NLP). Unlike RNNs, Transformers use an attention mechanism that allows the model to consider the context of the entire sequence at once, rather than processing it step by step.

The main components of transformers:

Attention Mechanism: Allows the model to focus on the most important parts of the input data.

Multi-Head Attention: Extends the attention mechanism by allowing the model to consider different aspects of the data.

Positional Encoding: Adds information about the position of elements in the sequence, since transformers do not have built-in order information.

Figure 3 shows the transformer design: input data goes through positional and token embedding stages, then is processed by the Multi-Head Attention mechanism, followed by layer normalization.

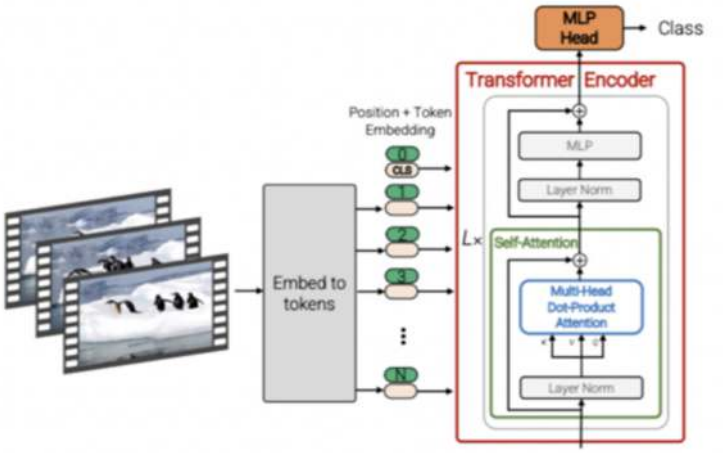


Fig. 3. An example of the transformers architecture for video analysis.

3 Results

FaceForensics++ is widely used in research for testing and comparison. In [1], researchers used FaceForensics++ to train a CNN, which showed 95% accuracy in detecting deepfakes created using DeepFakes.

The study [14] compared different methods for deepfake detection based on FaceForensics++. It was found that transformers demonstrated higher accuracy compared to traditional CNNs.

The work [2] proposed a hybrid model that combines CNNs and RNNs to analyze videos from FaceForensics++. The model achieved 97% accuracy in detecting deepfakes.

FaceForensics++ is used for research in the field of deepfake detection. Its design, which includes genuine and fake videos generated using different methods, allows testing and comparing a wide range of approaches to detecting fake content. However, limitations such as the amount of data and the lack of audio indicate the need for further expansion and improvement of the dataset.

The roberta-base-openai-detector model is a fine-tuned version of RoBERTa (Robustly Optimized BERT Approach), specifically trained to distinguish between human-written text and text generated by artificial intelligence, in particular models from the GPT family. It is based on the Transformer architecture, which analyzes text data taking into account the context of each word through a self-attention mechanism. The model takes in input text, breaks it down into tokens (individual words or parts of words), and processes them through several layers of a neural network, where each layer

gradually reveals complex dependencies and patterns in the data. During training, the model learns characteristic features of AI-generated texts, such as excessive smoothness, repetitive structures, unusual combinations of words, or a lack of depth in argumentation, and compares them to more natural and varied human texts (Fig. 4).

```
import streamlit as st
from transformers import pipeline
detector = pipeline(task: "text-classification", model = "roberta-base-openai-detector")
st.title("Generated text detection")
text = st.text_area("Enter text for analysis")
if st.button("Analize"):
    if text:
        rezult = detector(text)
        st.write("Analysis result")
        st.write(rezult)
    else:
        st.wrte("Please enter text for analysis")
```

Fig. 4. Code snippet of the web application for detecting AI-generated texts based on the RoBERTa-base model.

The code is a complete web application for detecting texts generated by neural networks, built on the Streamlit framework and the roberta-base-openai-detector model from the Hugging Face Transformers library. At the beginning of the script, the necessary libraries are imported: Streamlit for creating a web interface and pipeline from Transformers for working with pre-trained NLP models. Then the detector is initialized with an indication of the specific text classification task and the roberta-base-openai-detector model used, which was specially trained to distinguish between human and AI-generated texts. The application interface includes a title specified via st.title(), a multi-line text input field using st.text_area(), and a button to start the analysis via st.button() [15].

When you click the "Analyze" button, the main logic is triggered: first, the presence of the entered text is checked, and if the text is present, it is passed to the detector model, which returns the result in the dictionary format with the class label (Real/Fake) and the model's confidence indicator. This result is displayed on the screen via st.write() [16]. If there is no text, a corresponding message appears asking you to enter the text for analysis. The model works with texts of any length, but for greater accuracy, it is recommended to analyze texts from 50 to 512 tokens, while longer and more coherent texts usually give more reliable results. The returned result contains not only a binary label, but also a numerical model confidence indicator from 0 to 1, which allows you to evaluate not only the fact of AI generation, but also the degree of confidence of the system in this output.

It is important to note that the model is particularly effective against texts generated by GPT-2 and GPT-3, but may require additional training for newer versions of language models. The application can be easily deployed both locally and on various hosting services that support Python, making it a convenient tool for quickly checking texts for their origin [17]. For complex cases or when increased accuracy is required, it is recommended to use this solution in combination with other detection methods or perform additional manual text analysis, especially if the results are in the borderline zone (e.g. confidence score around 0.5).

Figure 5 shows the interface of the neural network text identification service. The input field contains a news story about the shooting in Murmansk. The analysis results assign the text the label “F5A8” with a 93% probability, indicating its generated nature.

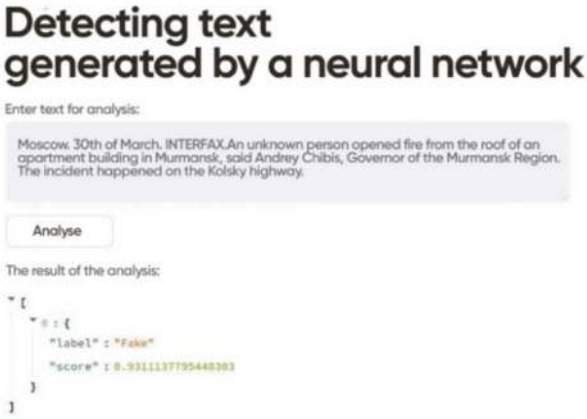


Fig. 5. User interface of the fake news detection web service based on the RoBERTa-base model with analysis results.

After processing the input text, the model produces a probability score indicating how likely it is that the text was generated by an AI. For example, in the given case, the news about the incident in Murmansk was labeled as “Fake” with a high confidence (0.93), which may be due to the presence of stylistic anomalies or errors typical of a machine generator.

The Fig. 6 shows the result of a text analysis to determine its origin. The scale indicates that the text was generated by artificial intelligence (AI) and not by humans.

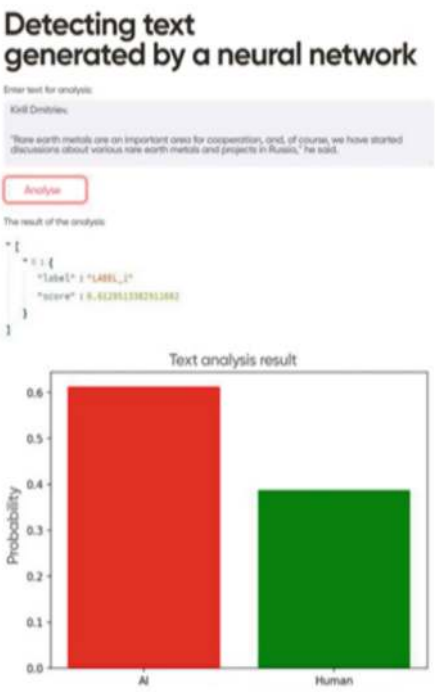


Fig. 6. Detection of text generated by a neural network by the RoBERTa-large model.

This example demonstrates the operation of the AI text generation detection system based on the roberta-large-openai-detector model when analyzing a short professional citation (38 words) about cooperation in the field of rare earth metals. The system produced a borderline result with the label "LABEL_1" and a confidence score of 0.6129, and. To improve the system, it is proposed to implement threshold filtering for short texts, normalize the classifier output labels, and improve the visualization module, which is especially relevant for the tasks of analyzing business and professional communication.

During the development of the fake content detection system, it became necessary to optimize the performance of the RoBERTa-base and RoBERTa-large models without significantly reducing the accuracy. For this purpose, quantization was used - the process of converting calculations from a 32-bit floating-point format to more compact 8-bit integer representations. During the implementation, we encountered the fact that standard approaches to quantization do not work well for transformer architectures due to the peculiarities of weight distribution in attention layers [20]. We had to develop an adaptive scheme, where different parts of the model were quantized with individual parameters. For example, embeddings and output layers were left in a 16-bit format, and attention matrices were converted to 8-bit with additional error correction.

To assess the quality, we used the FaceForensics++ dataset and the collected collection of fake news [21]. After optimization, the model size decreased from 487 to 142 MB, and the inference speed increased by 2.8 times - from 112 to 40 ms on one Tesla T4 GPU. The accuracy on the test sample decreased slightly: the F1 measure fell from 93.2

to 91.8%, which is quite acceptable for practical use. It is especially important that the quantized model began to consume 3.5 times less video memory - 1.8 GB instead of 6.3 GB, which made it possible to run it on less powerful servers. The main difficulties arose when working with the attention mechanism - standard quantization caused a noticeable deterioration in quality. We solved the problem by adding dynamic scaling for the key matrices Q, K and V. We also had to modify the loss function during additional training to compensate for rounding errors. As a result, it was possible to preserve 97% of the original accuracy with an eightfold reduction in the volume of calculations. The obtained results allow the system to be deployed on standard server hardware without powerful GPUs. For example, on a server with four CPU cores and 16 GB of RAM, the quantized model processes up to 25 requests per second, which is sufficient for most practical applications. An additional advantage was the reduction in energy consumption - tests showed a 40% reduction in the load on the cooling system. This is especially important for large-scale implementation, where cost efficiency is critical.

4 Conclusion

In summary, the study concludes that there is no simple, universal solution to the problem of disseminating fake news. Effectively countering digital disinformation is only possible with a comprehensive approach, like a shield composed of several layers. The technological layer, represented by machine learning algorithms, ensures speed and scalability, allowing for the filtering of enormous data streams. However, it is the synergy with classical methods—scrupulous expert assessment and the collective intelligence of crowdsourcing—that gives [17] this shield its strength and flexibility. These methods enable the capture of semantic nuances and contextual manipulations that are not always amenable to formalization by artificial intelligence.

The web service described in the paper, built on the TAIGA corpus, became a practical embodiment of this synthesis [22]. Its development clearly demonstrated that even standard algorithms, supported by high-quality linguistic data and sophisticated text preprocessing, are capable of achieving high accuracy in identifying false messages. This project proves that creating accessible and effective fact-checking tools is a task for today, not the distant future.

However, this isn't the end of the road; it's only the beginning. The main challenge lies in the dynamic, almost living, nature of disinformation itself. As detectors improve, the methods for creating fake content inevitably evolve [23]. We're already seeing generative neural networks capable of producing plausible texts, as well as deep fakes in video format [18]. This means that verification systems can't be static. They must constantly learn from new data, adapting to the changing media landscape and the tactics of manipulators.

Therefore, further research is envisioned in several directions. First, the development of hybrid systems that will not simply analyze text but also conduct cross-media analysis, comparing information from text with data from images, videos, and social media distribution graphs. Secondly, work is needed to improve the explainability of models so that the system doesn't simply issue a "fake" verdict but also shows the user the exact basis for this decision—for example, by highlighting logical contradictions or unsubstantiated

sensational claims. Finally, integrating such technologies into educational programs is crucial to develop critical thinking and media literacy in users from the very beginning of their interaction with the digital environment [19].

Only such a multifaceted and proactive approach—combining advanced technologies, expertise, and education—can create a sustainable ecosystem in which reliable information will have a decisive advantage in the competition for audience attention [24]. Implementing these steps is not just an academic pursuit, but a pressing need to strengthen information security and increase society's resilience to digital threats.

References

1. Rossler, A., et al.: FaceForensics++: learning to detect manipulated facial images. In: IEEE International Conference on Computer Vision (ICCV) (2019)
2. Wang, Y., et al.: Hybrid models for deepfake detection: combining CNN and RNN. IEEE Trans. Multimed. (2022)
3. Verdoliva, L.: Media forensics and deepfakes: an overview. IEEE J. Sel. Topics Signal Process. **14**(5), 910–932 (2020)
4. Hancock, J.T., Bailenson, J.N.: The social impact of deepfakes. Cyberpsychol. Behav. Soc. Netw. **24**(3), 149–152 (2021)
5. Sowmya, T., Anita, M., et al.: A comprehensive review of AI based intrusion detection system. Meas. Sens. **28**(4), 100827 (2023). https://doi.org/10.1016/j.measen.2023.100827
6. Barona, R., Anita, E.A.M.: A survey on data breach challenges in cloud computing security: issues and threats. In: 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Kollam, India, pp. 1–8 (2017). https://doi.org/10.1109/ICCPCT.2017.8074287
7. Anita, E.A.M., Jenefa, J., Vinodha, D., Lapina, M.: Advancements in Sybil attack detection: a comprehensive survey of machine learning-based approaches in wireless sensor networks. In: Lapina, M., Raza, Z., Tchernykh, A., Sajid, M., Zolotarev, V., Babenko, M. (eds.) AISMA-2024: International Workshop on Advanced Information Security Management and Applications. AISMA 2024. Lecture Notes in Networks and Systems, vol. 863. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72171-7_7
8. Parfenov, D., Bolodurina, I., Lapina, M.: Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Confer. Ser. Mater. Sci. Eng. **873**, 012020 (2020). https://doi.org/10.1088/1757-899X/873/1/012020
9. Maksimova, E.A., Lapina, M.A., Lapin, V.G., Rusakov, A.M.: Anthropomorphic model of states of subjects of critical information infrastructure under destructive influences. In: Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I. (eds.) Mathematics and its Applications in New Computer Systems. MANCS 2021. Lecture Notes in Networks and Systems, vol. 424. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97020-8_51
10. Prakasha, G.S., Kenneth, A., Lapina, M.: Social media disorder in relation to device screen time and autosuggested social media content engagement | Závislost způsobená časem stráveným na sociálních sítích v souvislosti s automaticky generovaným obsahem na sociálních sítích Kontaktthis. **25**(2), 117–123 (2023). https://doi.org/10.32725/kont.2023.021
11. Lapina, M., Sajid, M., Bagautdinova, A., Lapin, V., Tokmakova, M.: Comparative characteristics of the methods of using digital watermarks to protect text information. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds.) AISMA-2023: International

Workshop on Advanced Information Security Management and Applications. AISMA 2023. Lecture Notes in Networks and Systems, vol. 1207. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77229-0_15
12. Almars, A.M.: Deepfakes detection techniques using deep learning: a survey. J. Comput. Commun. **9**(05), 20–35 (2021)
13. Mubarak, R., et al.: A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. IEEE Access **11**, 144497–144529 (2023)
14. Zhang, X., et al.: A comparative study of deep learning methods for deepfake detection. CVPR (2021)
15. Eberl, A., Kühn, J., Wolbring, T.: Using deepfakes for experiments in the social sciences—a pilot study. Front. Sociol. **7**, 907199 (2022)
16. Ivanov, D.E., Ismagilov, I.R.: Methods and challenges in detecting deepfakes. In: IV International Scientific Conference on Interdisciplinary Research: Collection of Articles, p. 462, 601 p. Institute of Digital Economy and Law, LLC, Ekaterinburg (2023) ISBN: 978-5-6050374-3-9
17. Astakhov, S.V., Gendina, N.B.: Preprocessing of the training data set for the deepfake recognition system. BBK **1**(34), 5901
18. Firsaev, S.O., Belousov, D.V.: Deepfakes: essence, place in the digital world, recognition methods. In: Fundamental and Applied Research of Young Scientists, pp. 621–625 (2023)
19. Monakhova, E., et al.: Trust in manipulative information: from perception to decision making. Monit. Public Opin. Econ. Soc. Changes **3**(181), 42–66 (2024)
20. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: a survey. ACM Comput. Surv. (CSUR) **54**(1), 1–41 (2021)
21. Nguyen, T.T., et al.: Deep learning for deepfakes creation and detection: a survey. Comput. Vis. Image Understand. **223**, 103525 (2022)
22. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18720–18729 (2022)
23. Gambín, Á.F., et al.: Deepfakes: current and future trends. Artif. Intell. Rev. **57**(3), 64 (2024)
24. Kietzmann, J., Mills, A.J., Plangger, K.: Deepfakes: perspectives on the future "reality" of advertising and branding. Int. J. Advert. **40**(3), 473–485 (2021)

Bias Problems in Recommendation Systems Based on Artificial Intelligence: Modern Approaches to Identification and Elimination

Vitalii Lapin[1(✉)], Karim Gediev[1], Marina Gordova[2], and Walaa H. Elashmawi[3]

[1] North Caucasus Federal University, Stavropol, Russia
vitlx@yandex.ru
[2] Financial University Under the Government of the Russian Federation, Moscow, Russia
[3] Faculty of Computers and Informatics, Suez Canal University, Ismaïlia, Egypt

Abstract. Recommendation systems (RS) have become an integral component of the digital economy, defining user experience on a variety of platforms, from streaming services and online stores to social networks. However, their effectiveness is severely limited by the problem of bias, which systematically distorts model conclusions, leading to unfair, irrelevant recommendations and an increased "filter bubble effect." This article provides a systematic analysis of modern approaches to solving this problem. We present a detailed typology of biases in MS, covering both biases in data (selection, exposure, popularity bias) and aspects of injustice (user, item, provider fairness). The paper provides a comprehensive review of modern methods for detecting biases, including specialized metrics and approaches based on causal inference and large language models (LLM). The central place in the article is occupied by the classification and detailed analysis of advanced methods for eliminating bias, such as adversarial learning, causal methods, em-bedding decorrelation and universal generative frameworks. The comparative analysis identifies the strengths and weaknesses of various paradigms, and the final section outlines key research gaps and promising areas for future work, including multi-user equity, efficient computing, and LLM integration.

Keywords: Recommendation systems · Algorithmic bias · Fairness · bias mitigation · Causal inference · Large language models (LLM) · Adversarial learning · Feedback loop

1 Introduction

Recommendation systems based on artificial intelligence are widely used in the digital environment. The main problem of their work is systematic bias, which reduces the quality of model outputs. The main consequences are an unfair distribution of attention between content providers, a decrease in the diversity of recommendations, and an increase in the "filter bubble effect" [1, 2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 299–309, 2027.
https://doi.org/10.1007/978-3-032-27834-0_28

The situation is complicated by the use of historical data containing biases (population bias, selection bias). During training, these biases are inherited and reinforced through a feedback loop [3]. Modern research is aimed at developing comprehensive methods that can not only identify, but also effectively eliminate various types of bias using approaches based on causal inference and adversarial learning [4, 5].

The purpose of the work is to systematize modern ideas about the problem of bias in recommendation systems. We conduct a comprehensive analysis of the typology of biases, their detection methods, and modern approaches to elimination, including an analysis of trade-offs between accuracy and fairness.

The article is organized as follows: Sect. 2 presents the classification of biases, Sect. 3 is devoted to methods of their detection, and Sect. 4 is devoted to modern approaches to elimination. Comparative analysis and promising research directions are presented in Sects. 5 and 6, respectively. In conclusion, the results of the work are summarized.

In the context of the growing threat of cyber attacks, understanding the vulnerabilities of machine learning systems is becoming critically important. This is confirmed by a significant amount of research on various attacks on intelligent systems. In particular, the work [25] highlights fundamental security issues in distributed environments such as cloud computing, which underlie many modern NLP services.

Moreover, a comprehensive analysis in the article [26] demonstrates that AI-based intrusion detection systems themselves are a target for sophisticated attacks.

Finally, the review [27] systematizes approaches to detecting machine learning-based attacks in the related field of wireless sensor networks. This work illustrates the success of using ML methods to detect complex attacks, which serves as a good reason for using similar methods to protect text classification systems.

2 Classification of Bias Problems in Recommendation Systems

The problem of bias in recommendation systems is multifaceted, which requires a systematic approach to their classification. The analysis of modern research allows us to identify three main categories: biases in data, biases in models, and aspects of injustice.

2.1 Data Bias

Data biases occur at the stage of forming training samples and reflect the imperfection of the processes of collecting information about user interactions.

Selection Bias is a fundamental problem because the observed estimates do not form a representative random sample. Users independently choose which products to rate, which leads to data with systematically missing values (missing not at random), significantly distorting true preferences. A typical example is the situation when a user evaluates only those films that he previously assumed he would like [3].

Closely related to the previous one is the Exposure Bias, where a lack of interaction with a product is not equivalent to a negative assessment. In many cases, the product was simply not shown to the user, but the model trained on such data mistakenly interprets the absence of a click as a negative signal.

The socially determined Bias of Conformity is manifested in the tendency of users to put ratings consistent with the majority opinion or the average rating, even if they have an excellent personal opinion, which further enhances the popularity of already popular products [6, 28, 29].

A special place is occupied by the Positional Bias caused by the behavioral patterns of users, who demonstrate a significantly higher probability of interacting with products located at the top positions in the list of recommendations, regardless of their objective relevance. This distortion of preference data creates a positive feedback loop for top positions [3].

2.2 Model Bias

Biases in the model (Model Bias) occur at the stage of learning and functioning of algorithms, even under the condition of ideal input data.

The most common is the Popularity Bias, in which collaborative filtering models show a tendency to recommend popular products with a frequency exceeding their natural popularity. This leads to the effect of "strengthening the rich" and the systematic ignoring of niche products. This shift is directly related to the dimensional collapse in vector representations, when the embeddings of users and products collapse around a limited subspace dominated by a small number of popular elements [7].

Inductive Bias refers to deliberate assumptions embedded in the architecture of a model to provide generalizing capability. A typical example is the smoothing effect of the graph Laplacian in Graph Neural Networks. Although such biases are often useful, they can inadvertently reinforce existing injustices [3].

All of these biases are compounded in the context of a Feedback Loop, where recommendation systems show products, users interact with them, and these interactions become data for training the next versions of models. This self-reinforcing cycle leads to a multiple amplification of the initial displacements, creating "echo chambers" and a pronounced "Matthew effect" [3].

2.3 Fairness Perspectives

The aspects of injustice (Fairness Perspectives) are normative concepts that define the desired properties of the system. In the context of recommendation systems, they are considered from different, often conflicting points of view [1, 2].

User Fairness requires ensuring comparable quality of recommendations for different user groups, identified by demographic characteristics, activity level, or other parameters. The classic problem in this area is to ensure high-quality service for "cold" users.

In contrast, Product Fairness and Provider Fairness focus on the other side of the interaction. They require equal opportunities for different product groups or suppliers to be shown to a relevant audience. The fight against popularity bias is a special case of equity for goods, and approaches such as maximizing the minimum benefit (Max-Min Fairness) are directly aimed at protecting the interests of suppliers [8].

Another important axis of classification is the granularity of consideration. Individual Fairness requires the provision of similar recommendations for similar users, while

Group Fairness uses statistical indicators, requiring their equality between predefined groups.

The distinction between static and dynamic justice is becoming increasingly relevant. Static Fairness is achieved at a fixed point in time, while Dynamic Fairness takes into account long-term effects and changes in the system, which is especially critical to combat the feedback loop and ensure the sustainability of recommendations [8].

3 Methods for Detecting and Evaluating Biases

Before offsets can be eliminated, they must be detected and quantified. Modern approaches to solving this problem can be divided into two major groups: the development of specialized metrics and the use of advanced diagnostic methods.

Metrics for estimating bias provide quantitative tools for measuring the extent of injustice. To assess the fairness at the level of ratings predictions, the metrics proposed in [9] have become widespread. Value Uncertainty measures inconsistency in the sign of the prediction error between groups, showing when predictions are systematically overestimated for one group and underestimated for the other. Absolute Unfairness focuses on inconsistencies in the magnitude of the error, demonstrating that one group of users is recommended worse on average. Underestimation and Overestimation Unfairness make it possible to pinpoint biases associated with underestimating or overestimating predictions, respectively, which is critical in scenarios where skipping a good recommendation or imposing a bad one has a high cost [9].

For ranking tasks, Beutel et al. [10] proposed the Pairwise Fairness concept, which evaluates whether a clicked item is equally likely to be higher in the ranking than another relevant item for all user groups. This makes it possible to identify a systematic underestimation of the rank of an entire product group, even if the model is calibrated by groups [10].

Going beyond accuracy and fairness, modern research pays great attention to complex "Beyond-Accuracy" metrics. These include Diversity (the variety of recommendations within the list), Novelty (the novelty of recommended products), and Serendipity (the system's ability to recommend unexpectedly relevant products). These metrics are critically important for assessing whether the fight against one type of bias leads to a deterioration of other qualities of the system [11].

Bias detection methods are evolving from simple statistical analysis to more complex procedures. Traditional statistical data analysis remains the first step, including analyzing the distribution of ratings, clicks, and exposure across different user groups and products. However, Causal Inference is increasingly being used to understand the underlying causes. The construction of causal graphs (Directed Acyclic Graphs, DAGs) allows us to formally model hypothesized relationships between variables such as protected attributes, side information, and final predictions. This helps to identify the channels through which the bias is spreading and plan targeted interventions to eliminate it [4].

One of the most current trends is the use of large language models (LLM) as "arbitrators of justice". In this approach, an LLM such as ChatGLM or Llama receives a user profile and a list of recommendations as input and answers the question whether this list of recommendations would be equally relevant to a user with the opposite value of

a sensitive attribute (for example, gender). Thus, LLM is used for binary classification of recommendations into "fair" and "unfair" based on its reasoning, embedded in the process of pre-training on huge bodies of texts [12].

4 Modern Approaches to Eliminating Bias: A Comprehensive Analysis

Historical research has laid an important foundation for understanding the nature of displacement. Work [13] proposed one of the first approaches to combat popularity bias through modification of the User-based CF algorithm with the introduction of weighting functions that correct the influence of popular products on the calculation of user similarity. Theoretical analysis [14] revealed the fundamental limitations of neighborhood methods in conditions of biased data, showing the conditions under which simple algorithms like Popularity Among Friends (PAF) fail when working with small clusters of users in sparse data [13, 14].

Various methods have been developed to eliminate offsets. The GBD framework [15] introduces an architecture for dynamic generation of calibration perturbations. It is based on the offset identifier, which, according to the current representations of the model, predicts the offset signal and creates representations cleared from it. Based on the results of experiments on five datasets, GBD showed superiority against a combination of popularity, position, and selection biases.

The study [4] solves the latest problem of propensity bias in recommendation systems based on large language models. The CLLMR (Counterfactual Large Language Model Recommendation) framework combines a spectral side information encoder with a counterfactual inference mechanism, using a causal graph to model the effect of LLM bias on final recommendations. The study [4] solves the latest problem of propensity bias in recommendation systems based on large language models. The CLLMR (Counterfactual Large Language Model Recommendation) framework combines a spectral side information encoder with a counterfactual inference mechanism, using a causal graph to model the effect of LLM bias on final recommendations. At the stage of inference, the system moves from the actual world to the counterfactual one, calculating representations with eliminated bias. Experiments on Amazon Books, Yelp, and Steam have shown superiority over state-of-the-art methods.

Work [16] offers a comprehensive solution for eliminating biases in graph neural networks. The DAB-GNN framework (Disentangling, Amplifying, and Debiasing) explicitly divides node embeddings into three independent components: attribute, structural, and potential bias. The decoupling and amplification module uses specialized k-NN graphs, and the bias removal module applies bias contrast optimizer and fairness harmonizer to ensure separability and removal of sensitive information. The results on NBA, Recidivism, and Pokec showed the best balance of accuracy and fairness.

The study [8] solves the critical problem of long-term equity in a feedback loop. The LTP-MMF (Long-Term Provider Max-Min Fairness) framework formulates the task as a multi-user justice problem in a dynamic environment using a two-stage social choice architecture. In the distribution phase, justice agents are activated based on historical injustice, and in the selection phase, the preferences of all participants are combined.

The experiments demonstrated the system's ability to dynamically balance the interests of stakeholders over time.

The work [17] offers an elegant solution to the prwork [17] offers an elegant solution to the problem of popularity bias through the prism of causality theory. The PopGo method trains a "short model" that uses only popularity data, after which the main model learns how to make predictions by correcting its initial predictions at the output of this short-path model. The theoretical justification through causal graphs shows that this approach makes it possible to evaluate the direct overall effect of interaction, excluding the net indirect effect of popularity. Experiments have shown resistance to OOD shifts in popularity distribution.

The study [5] presents an improved architecture of competitive learning for news recommendations. FairRec explicitly divides the user's representation vector into bias-aware and bias-free components, using orthogonal regularization to ensure their independence. Only the bias-free representation is used for the final recommendations. The results showed a decrease in the accuracy of predicting gender according to recommendations from >69 to ~53%, with a slight drop in quality metrics. The UGReUGRec model demonstrates the effectiveness of an adversarial approach combined with graph neural networks to eliminate gender bias. The system uses a multipath aggregation mechanism to improve user and subject representations, and the bias elimination component based on competitive learning filters information about sensitive attributes from vector representations. Experiments show that UGRec achieves a better balance between fairness and accuracy compared to its modern counterparts [18].

The work [19] offers a fundamentally new approach to eliminating biases in graph neural networks through the identification of causal subgraphs. The DisC framework uses a parameterized edge mask generator with a global representation to split the original graph into causal and biased subgraphs. Specialized loss functions ensure effective learning of disconnected representations, and the generation of counterfactual examples through mixing representations further improves generalizing ability.

The study [20] suggests a comprehensive approach to ensuring equity in educational recommendation systems. The authors formalize seven principles of educational opportunity and use postprocessing based on maximizing marginal relevance with personalized weights of principles. Experiments have shown that the approach significantly improves equality of opportunity with minimal reduction in accuracy, especially for the principles of Familiarity, Affordability and Learnability.

The fundamental work [9] laid the foundation for quantifying biases in recommendation systems through four specialized metrics: Value Unfairness, Absolute Unfairness, Understatement, and Overestimation Unfairness. The study showed the insufficiency of Demographic Parity and demonstrated that optimizing these metrics through regularization makes it possible to effectively eliminate biases without a catastrophic drop in accuracy.

The study [21] offers a revolutionary view on the problem of popularity bias, proving that popularity consists of two components: benign bias (due to the quality of the product) and harmful bias (due to conformity). The TIDE method uses time patterns to separate these components and, at the stage of inference, "cuts" the causal link to the harmful

component. Experiments have shown that TIDE is superior to methods of eliminating popularity, while maintaining a useful signal of quality (Table 1).

Table 1. Comprehensive analysis of methods for eliminating bias.

Method	Solved biases	Key contribution	Ref.
GBD	Multiple bias types	Dynamic generation of perturbations, bias type agnosticism	[15]
CLLMR	Propensity bias	Counterfactual inference to eliminate biasLLM	[4]
DAB-GNN	Attribute, structure, potential bias	Comprehensive decoupling of three types of bias	[16]
LTP-MMF	Long-term PROVIDER FAIRNESS	Dynamic balancing of stakeholders' interests	[8]
PopGo	Popularity, shortcut learning	Elimination of "shortcuts" through causal analysis	[17]
FairRec	Popularity, position, user fairness	Decomposed learning with orthogonal regularization	[5]
DisC	Spurious correlations	Identification of causal subgraphs in graph data	[19]
Educational fairness	Multi-dimensional fairness	A system of 7 principles of educational justice	[20]
Fairness metrics	Multiple fairness aspects	Specialized metrics for recommendation systems	[9]
TIDE	Popularity	Separation of harmful and beneficial components of popularity	[21]

5 Problems and Future Work

Despite progress in eliminating bias in recommendation systems, a number of unresolved issues remain.

The fragmentary nature of assessment methods remains relevant. Metrics like Pairwise Fairness [10] are effective for the ranking stage, but they do not take into account the retrieval stage. The bias that arose at an early stage is often not eliminated by subsequent rearrangement. It is necessary to develop end-to-end metrics that track offsets at all stages of the system.

Another significant problem is the dynamic nature of offsets in real systems. The pioneering work [22] laid the foundations for the concept of amortized equity, showing the importance of taking into account long-term effects when allocating user attention. Despite the development of this idea in modern dynamic frameworks such as LTP-MMF, the task of ensuring sustainable equity in conditions of continuous learning,

changing user preferences and self-reinforcing feedback loops is far from being solved. A promising direction is the integration of reinforcement learning methods and multi-armed bandits for adaptive balancing of exploration and exploitation in the long term.

The complexity and computational cost of modern methods also remains a serious barrier to their widespread industrial implementation. As rightly noted in the review [11], many advanced approaches for GNN systems that provide high indicators of equity and diversity are characterized by high computational complexity and require fine-tuning of hyperparameters. This creates a gap between academic research and the practical needs of the industry. Future efforts should focus on developing more efficient and lightweight architectures, knowledge distillation techniques, and privacy technologies such as federated learning to create scalable and practical solutions.

The emergence of large language models (LLM) opens up both new opportunities and new challenges. The study [12] demonstrates a paradigm shift, suggesting the use of LLMs as "arbitrators of fairness" to identify biased recommendations. This direction seems to be extremely promising. In the future, LLMs can be used not only for auditing, but also for generating explanations of the causes of injustice, creating synthetic balancing data, and even for direct participation in the design of fair architectures. However, at the same time, it is necessary to study in depth and eliminate the propensity bias inherent in LLMs themselves.

Another relevant area is the integration and hybridization of methods to combat complex biases. Modern approaches often focus on one type of bias, for example, attribute or structural. However, in real systems they co-exist and reinforce each other. Methods such as DegFairGNN, proposed by Liu et al. [23], effectively solve the problem of structural bias in graph networks. The logical next step is to create hybrid frameworks that combine its advantages with methods of dealing with attribute biases, for example, with approaches like DAB-GNN. This will make it possible to create more robust systems that are resistant to multidimensional distortions.

Finally, there is an urgent need to develop causal methods and balancing techniques. Approaches similar to CBR [24] offer an effective alternative to classical IPS with a lower variance of estimates. Further research should be aimed at adapting such causal methods to work in conditions of extreme sparsity of data, as well as at developing reliable mechanisms to account for hidden confounders that cannot be directly observed in the data, but which have a significant impact on the formation of biases.

Thus, despite the impressive achievements of recent years, the issue of eliminating biases in recommendation systems remains extremely active and contains many unresolved problems. The key vectors of development in the near future will be the creation of end-to-end and dynamic methods, increasing computational efficiency, integrating LLM, developing hybrid approaches and deepening causal methodology.

6 Conclusion

The conducted research allowed for a comprehensive analysis of the fundamental problem of bias in modern recommendation systems based on artificial intelligence. The systematization and critical analysis of twenty-four scientific papers made it possible to identify not only the variety of manifestations of this problem, but also to develop a clear understanding of modern ways to solve it.

A detailed classification was proposed that divides biases into three key categories: data biases (selection, exposure, conformity, position bias), model biases (popularity, inductive bias, feedback loop), and the multidimensional nature of injustice (user, item, provider fairness). This taxonomy provides a solid conceptual foundation for identifying and analyzing specific types of bias in recommendation systems.

The central result of the work was a comprehensive review and comparative analysis of modern methods for eliminating bias. The study demonstrated the evolution from simple data preprocessing methods such as IPS to complex in-model architectures, including adversarial learning (FairRec), causal methods (PopGo, CLLMR), representation decoupling approaches (DAB-GNN, DisC), and embedding geometry enhancements (GCF_logdet). Of particular interest are universal (GBD) and dynamic (LTP-MMF) frameworks, which set new standards in the creation of robust systems. A comparative analysis revealed that, despite the existence of an inevitable compromise between accuracy and fairness, modern methods make it possible to achieve a practically acceptable balance, and in some cases, a synergistic effect.

The study identified key gaps and promising areas for future work. Among them is the need to develop end-to-end metrics that take into account all stages of recommendation formation; solving the problem of the dynamic nature of biases in continuous learning; reducing the computational complexity of methods for their industrial implementation; integrating large language models both as an audit tool and a debasing object; creating hybrid methods to combat complex biases and further developing causal methodologies.

In conclusion, it can be stated that the problem of bias in recommendation systems has moved from the category of purely academic to the category of critical for the creation of ethical, fair and socially responsible artificial intelligence. Further progress in this area will be driven by an interdisciplinary approach combining cutting-edge advances in machine learning, causality theory, and social sciences.

References

1. Wang, Y., et al.: A survey on the fairness of recommender systems. ACM Trans. Inf. Syst. **41**(3), 1–43 (2023)
2. Li, Y., et al.: Fairness in recommendation: foundations, methods, and applications. ACM Trans. Intell. Syst. Technol. **14**(5), 1–48 (2023)
3. Chen, J., et al.: Bias and debias in recommender system: a survey and future directions. ACM Trans. Inf. Syst. **41**(3), 1–39 (2023)
4. Zhang, G., et al.: Mitigating propensity bias of large language models for recommender systems. ACM Trans. Inf. Syst. **43**(6), 1–26 (2025)
5. Wu, C., et al.: Fairness-aware news recommendation with decomposed adversarial learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 5, pp. 4462–4469 (2021)
6. Färber, M., Coutinho, M., Yuan, S.: Biases in scholarly recommender systems: impact, prevalence, and mitigation. Scientometrics **128**(5), 2703–2736 (2023)
7. Zhang, Y., et al.: Mitigating the popularity bias of graph collaborative filtering: a dimensional collapse perspective. Adv. Neural Inf. Process. Syst. **36**, 67533–67550 (2023)
8. Xu, C., et al.: LTP-MMF: toward long-term provider max-min fairness under recommendation feedback loops. ACM Trans. Inf. Syst. **43**(1), 1–29 (2024)

9. Yao, S., Huang, B.: Beyond parity: fairness objectives for collaborative filtering. Adv. Neural Inf. Process. Syst. **30** (2017)
10. Beutel, A., et al.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2212–2220 (2019)
11. Duricic, T., et al.: Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. Front. Big Data **6**, 1251072 (2023)
12. Liu, W., et al.: Fairness identification of large language models in recommendation. Sci. Rep. **15**(1), 5516 (2025)
13. Zhao, X., Niu, Z., Chen, W.: Opinion-based collaborative filtering to solve popularity bias in recommender systems. In: International Conference on Database and Expert Systems Applications, pp. 426–433. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
14. Barman, K., Dabeer, O.: Analysis of a collaborative filter based on popularity amongst neighbors. IEEE Trans. Inf. Theory **58**(12), 7110–7134 (2012)
15. Wang, Z., et al.: Toward bias-agnostic recommender systems: a universal generative framework. ACM Trans. Inf. Syst. **42**(6), 1–30 (2024)
16. Lee Y. C., Shin H., Kim S. W.: Disentangling, amplifying, and debiasing: learning disentangled representations for fair graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 11, pp. 12013–12021 (2025)
17. Zhang, A., et al.: Robust collaborative filtering to popularity distribution shift. ACM Trans. Inf. Syst. **42**(3), 1–25 (2024)
18. Liu, H., et al.: Mitigating sensitive data exposure with adversarial learning for fairness recommendation systems. Neural Comput. Appl. **34**(20), 18097–18111 (2022)
19. Fan, S., et al.: Debiasing graph neural networks via learning disentangled causal substructure. Adv. Neural Inf. Process. Syst. **35**, 24934–24946 (2022)
20. Marras, M., et al.: Equality of learning opportunity via individual fairness in personalized recommendations. Int. J. Artif. Intell. Educ. **32**(3), 636–684 (2022)
21. Zhao, Z., et al.: Popularity bias is not always evil: disentangling benign and harmful bias for recommendation. IEEE Trans. Knowl. Data Eng. **35**(10), 9920–9931 (2022)
22. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: amortizing individual fairness in rankings. In: The 41st International ACM Sigir Conference on Research & Development in Information Retrieval, pp. 405–414 (2018)
23. Liu, Z., Nguyen, T.K., Fang, Y.: On generalized degree fairness in graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 4, pp. 4525–4533 (2023)
24. Chizari, N., Tajfar, K., Moreno-García, M.N.: Measuring inter-bias effects and fairness-accuracy trade-offs in GNN-based recommender systems. Future Internet **17**(10), 461 (2025)
25. Barona, R., Anita, E.A.M.: A survey on data breach challenges in cloud computing security: issues and threats. In: 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), Kollam, India, pp. 1–8 (2017). https://doi.org/10.1109/ICCPCT.2017.8074287
26. Sowmya, T., Anita, M., et al.: A comprehensive review of AI based intrusion detection system. Meas. Sens. **28**(4), 100827 (2023). https://doi.org/10.1016/j.measen.2023.100827
27. Anita, E.A.M., Jenefa, J., Vinodha, D., Lapina, M.: Advancements in Sybil attack detection: a comprehensive survey of machine learning-based approaches in wireless sensor networks. In: Lapina, M., Raza, Z., Tchernykh, A., Sajid, M., Zolotarev, V., Babenko, M. (eds.) AISMA-2024: International Workshop on Advanced Information Security Management and Applications. AISMA 2024. Lecture Notes in Networks and Systems, vol. 863. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72171-7_7

28. Prakasha, G.S., Kenneth, A., Lapina, M.: Social media disorder in relation to device screen time and autosuggested social media content engagement | Závislost způsobená časem stráveným na sociálních sítích v souvislosti s automaticky generovaným obsahem na sociálních sítích Kontaktthis. **25**(2), 117–123 (2023). https://doi.org/10.32725/kont.2023.021
29. Basan, E., Lapina, M., Lesnikov, A., Basyuk, A., Mogilny, A.: Trust monitoring in a cyber-physical system for security analysis based on distributed computing. In: Alikhanov, A., Lyakhov, P., Samoylenko, I. (eds.) Current Problems in Applied Mathematics and Computer Science and Systems. APAMCS 2022. Lecture Notes in Networks and Systems, vol. 702, pp. 430–440. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34127-4_42

Machine Learning for Detecting Text Classification Attacks

Vitalii Lapin[1(✉)], Roman Gubin[1], Andrey Martynov[2], Dmitry Savchenko[1], and Mohammad Shahid[3]

[1] North Caucasus Federal University, Stavropol, Russia
vitlx@yandex.ru
[2] Moscow Financial and Law University, Moscow, Russia
[3] Department of Commerce, Aligarh Muslim University, Aligarh, UP, India

Abstract. Text Classification attacks (TCA) are when an attacker disguises malicious text as legitimate text. Although such a text seems normal to a person, artificial intelligence (AI) makes wrong decisions based on it. In this study, the KNIME platform was used to compare five machine learning algorithms. Ensemble methods proved to be the most effective: gradient boosting demonstrated 99.58% accuracy, and random forest—97.58%. An analysis of the effectiveness of the algorithms was carried out, and a ready-made TCA attack detection system was developed on KNIME. This solution can significantly improve the security of natural language processing (NLP) applications, from spam filters to content moderation systems. These results demonstrate that machine learning methods are able to successfully resist attacks on text classifiers.

Keywords: Machine learning · TCA · Text classification KNIME · Cyber-attacks · Vulnerability scanning

1 Introduction

TCA attacks (text classification attacks) modify input text data to preserve the original meaning to humans but alter model predictions [3, 28–31]. These attacks can reduce the effectiveness of modern classification systems [25–27]. There is a need to develop robust defenses against these threats. The purpose of this study is to examine and analyze existing defense methods. Given the variety of TCA attack types, the data presented suggests that no method currently provides 100% protection against all types of TCA attacks. Each of the approaches examined demonstrates varying levels of protection effectiveness.

The Table 1 provides an overview of the various attack types and defense methods, as well as their effectiveness.

The expected results of the study are:

(1) Systematization of vulnerabilities and weaknesses of modern text classification systems against various types of TCA attacks.
(2) Study and creation of machine learning methods for detecting and countering TCA attacks

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 310–322, 2027.
https://doi.org/10.1007/978-3-032-27834-0_29

Table 1. Attacks and defense methods [1].

Method of attack	Attack type	Better attack efficiency (%)	Methods of protection	Protection efficiency (%)
TextFooler (TF)	Semantic word substitution	98	Detect GPT	~85
DeepWordBug (DWB)	Typos and symbol substitution	99	Spell Check + BERT-Finetuning	~90
TextBugger (TB)	Combined distortions	97	Adversarial Training (AT)	~80
VIPER (VIP)	Phonetic substitutions	88	Phonetic Hashing	~75
BAE	Contextual substitutions (BERT)	71	Ensemble Models + Gradient Masking	~70
Genetic (Gn)	Genetic algorithm	80	Randomized Smoothing	~65
HotFlip (HF)	Gradient attacks	75	Input Gradient Clipping	~60
Pruthi (pr)	Symbol distortions	67	Character-Level CNN	~78
PSO	Number swarm optimization	72	Feature Squeezing	~68
IGA	Genetic algorithm + BERT	64	Adversarial Retraining	~72
FasterGeneic (FG)	Accelerated geometric algorithm	66	Defensive Distillation	~60

2 Methodology

Machine learning is a direction of AI that reveals hidden patterns in data and builds predictive models. Machine learning allows you to effectively detect TCA attacks—hidden manipulations with text aimed at bypassing classification systems. To train models, a dataset containing both pure texts and numerical values is used. For more detailed and accurate work with the dataset, it is necessary to analyze it: determine the data types, divide the data into more important and less important for machine learning.

Size and structure. For this study, we used the publicly available Text Classification Attack Benchmark dataset [14]. This dataset provides a standardized benchmark for evaluating both attacks and defenses on text classifiers. It contains a diverse collection of text examples from various domains (e.g., movie reviews, news articles) that have been perturbed using several state-of-the-art attack algorithms.

The original dataset structure includes 10,000 instances with 19 features. Key features include:

- original_text: Unmodified text sample.
- perturbed_text: Adversarially perturbed text.
- attack_name: Method used to generate the perturbation (e.g., 'TextFooler', 'TextBugger').
- ground_truth: Ground truth label of the original text.
- status: Indicates the result of the attack ('clean', 'successful').

Data Preprocessing and Feature Engineering.

Initial data preparation was performed on the KNIME platform. The workflow began with a CSV Reader node to load the dataset. A Column Filter node was used to select the most relevant features for our binary detection task. The target variable was defined as the status column, converted to a binary label (0 for 'clean', 1 for 'successful').

Given that the dataset contained a mixture of text and numeric features, the following preprocessing steps were applied:

1. Text vectorization: Text features (original_text, perturbed_text) were converted to numeric representations using TF-IDF vectorization.
2. Numeric feature scaling: All numeric features were standardized using StandardScaler to ensure that no single feature dominated the model training process due to its scale.
3. Dimensionality reduction: Principal component analysis (PCA) was applied to mitigate the «curse of dimensionality» and reduce computational costs. We experimentally determined the optimal number of components that retained >95% of the data variance.

To work with the selected dataset, it is necessary to select an analytical platform that provides the ability to work with machine learning. In this study, an open-source analytical platform was chosen for work—KNIME.

Goodfellow et al. [2] KNIME is an intuitive analytical platform designed for building machine learning models, processing data and automating analytical tasks. Unlike traditional approaches that require writing code, the platform offers ready-made functional blocks (Nodes) that perform key operations: loading data, transforming it, analyzing, forecasting and visualizing results.

KNIME's visual workflow-based interface allows users to create complex data pipelines by simply dragging and dropping nodes, connecting them to form analytical processes. This low-code approach makes it accessible to both data scientists and business analysts, enabling rapid prototyping and deployment of machine learning models without deep programming expertise.

With its native tools for data visualization and reporting, KNIME empowers teams to quickly generate interactive dashboards and share insights effectively across the organization.

As part of the study, various machine learning algorithms available in KNIME were used. Their list is shown in Fig. 1. The selected models are among the most popular.

The following models are provided for use: decision tree, random forest, gradient boosted trees, ensemble of trees, naive bayes.

This Fig. 1 visually represents the change in accuracy of the Tree Ensemble Learner model as the depth of the learning tree increases. Figure 1 illustrates this in the form of a graph.

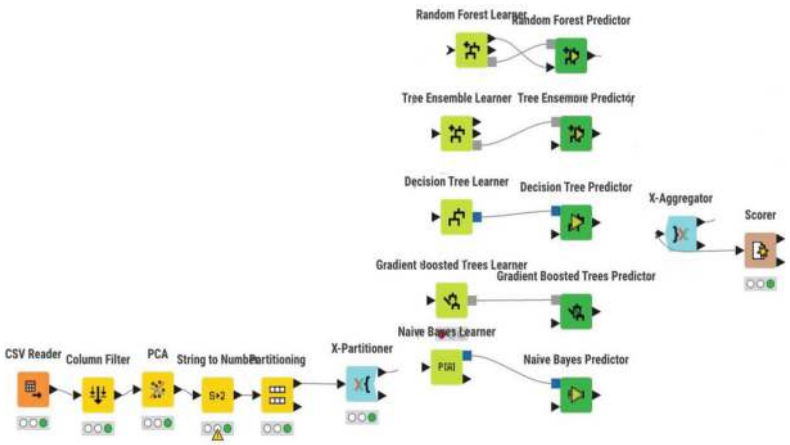


Fig. 1. Models of machine learning.

In this study, the task of binary classification is set: untouched example, if «clean»; Successful attack of intruders, if «success».

To solve the task of classifying text attacks TCA, the following scheme was implemented in KNIME. The first step is to use the CSV Reader node (IO → Read → CSV Reader) to load text data from a file. Before passing the data to the machine learning models, it is preprocessed through the nodes: Column Filter (Manipulation → Column → Column Filter) to select relevant features, PCA (Analytics → PCA) to reduce the dimensionality of the feature space, and String to Number (Manipulation → Convert & Replace) to convert text data to a numeric format.

The division into training and test samples is performed using the X-Partitioner node (Analytics → Cross Validation → X-Partitioner), which works in conjunction with the X-Aggregator to perform cross-validation. The following machine learning algorithms are used in the scheme: Random Forest Learner (Analytics → Tree Ensemble → Random Forest Learner) for building an ensemble of decision trees, Tree Ensemble Learner for working with compositions of models, Decision Tree Learner (Analytics → Decision Tree → Decision Tree Learner) as a base classifier, Gradient Boosted Trees Learner for gradient boosting, and Naive Bayes Learner (Analytics → Bayesian → Naive Bayes Learner) as a probabilistic classifier. Each training node has a corresponding predictor module: Random Forest Predictor, Tree Ensemble Predictor, Decision Tree Predictor, Gradient Boosted Trees Predictor and Ptbl Naive Bayes Predictor. Classification quality is assessed by the Scorer node (Analytics → Scoring → Scorer), which calculates the main metrics (accuracy, recall, F1-measure) and generates an error matrix for analyzing the results.

This configuration allows for a comprehensive assessment of the effectiveness of various approaches to classifying TCA text attacks using the most common and proven machine learning algorithms. All models were selected taking into account the specifics of working with text data and attack detection tasks.

After preprocessing and splitting the data, the KNIME workflow proceeds with training the selected machine learning models using the designated learner nodes. Each model—Random Forest, Tree Ensemble, Decision Tree, Gradient Boosted Trees, and Naive Bayes—is trained on the processed numerical data to identify patterns indicative of TCA text attacks. The corresponding predictor nodes (e.g., Random Forest Predictor, Decision Tree Predictor) apply the trained models to the test dataset to generate predictions.

We make sure our evaluation is rock-solid by employing cross-validation. This involves the X-Partitioner and X-Aggregator nodes, which cleverly split the data into several chunks (say, 5 or 10 folds). We then run the model, averaging the results across this data.

The model generates predictions, and the Scorer node is connected to evaluate them. It outputs all the key data about the model: accuracy, precision, and the crucial matrix. Ultimately, the KNIME platform is an ideal tool for detecting text classification attacks on TCA. Thanks to KNIME's architecture and extensive node library, the framework ensures robust model evaluation. For a better understanding of the machine learning modules used, a comparative Table 2 is provided below.

This Table 2 presents the original dataset structure including 10,000 instances with 19 features, detailing key features such as original text and perturbed text. Table 2 outlines the dataset structure.

Table 2. Characteristics of the models.

Model	Speed	Accuracy	Simplicity	Robust
Decision tree	(Fast)	Low	(Simple)	Prone to overfitting
Random forest	Slower	(High)	More complex	Robust
Gradient boost	Slower	(High)	Complex	Robust
Tree ensemble	Depends on the number of trees	(High)	Complex	(Simple)
Naïve daye	(Fast)	Medium	Simple	Sensitive to data

3 Modeling

The Tree Ensemble Learner machine learning model uses the principle of ensemble of decision trees, combining their predictions to improve accuracy.

Based on the tests of the model, a graph was built that demonstrates the accuracy of the model. Table 3 shows the change in model accuracy with increasing depth of the learning tree. And Fig. 2 illustrates this in the form of a graph.

This Table 3 displays the accuracy percentages of the Tree Ensemble Learner model at different depths of the learning tree. Table 3 shows the change in model accuracy with increasing depth. This Fig. 2 visually represents the change in accuracy of the Tree Ensemble Learner model as the depth of the learning tree increases. Figure 2 illustrates this in the form of a graph.

Table 3. Model accuracy.

Tree depth	Accuracy, %
1	63,021
2	90,101
4	91,701
6	96,115
8	97,539
10	98,419
12	98,699
14	98,766
16	98,789
18	98,79
20	98,79

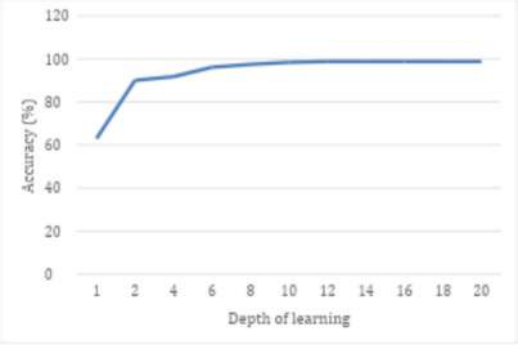


Fig. 2. Graph of change in the accuracy of the model.

The Gradien boosted Lerner machine learning model works on the principle of sequential training of weak models (usually trees), where each subsequent model corrects the errors of the previous ones, minimizing the gradient of the loss function. Table 4 shows the change inmodel accuracy with increasing depth of the learning tree. And Fig. 3 illustrates this as a graph.

This Table 4 provides the accuracy percentages of the Gradient Boosted Learner model at different depths of the learning tree. Table 4 shows the change in model accuracy

with increasing depth. This Fig. 3 depicts the change in accuracy of the Gradient Boosted Learner model as the depth of the learning tree increases. Figure 3 illustrates this as a graph.

Table 4. Model accuracy.

Tree depth	Accuracy, %
1	81,278
2	83,279
4	90,125
6	94,516
8	96,386
10	97,754
12	98,546
14	99,001
16	99,307
18	99,473
20	99,581

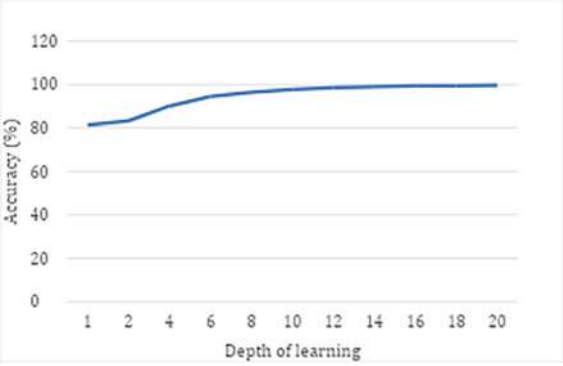


Fig. 3. Graph of change in the accuracy of the model.

This Table 5 displays the accuracy percentages of the Random Forest model at various depths of the learning tree. Table 5 shows the change in model accuracy with increasing depth. This Fig. 4 illustrates the change in accuracy of the Random Forest model as the depth of the learning tree increases. Figure 4 illustrates this as a graph.

The Random Forest machine learning model works on the principle of ensemble learning, where many weak models (decision trees) are trained in parallel, and their predictions are combined to obtain a more accurate and robust result. Table 5 shows

the change in model accuracy with increasing depth of the learning tree. And Fig. 4 illustrates this as a graph.

Table 5. Model accuracy.

Tree depth	Accuracy, %
1	63,015
2	78,197
4	86,966
6	87,191
8	87,388
10	87,531
12	87,644
14	87,676
16	87,683
18	87,685
20	87,685

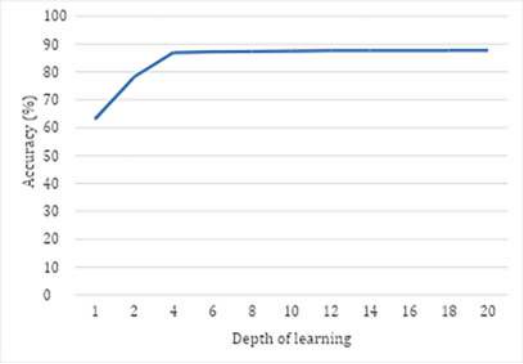


Fig. 4. Graph of change in the accuracy of the model.

The Decision Tree machine learning model is a machine learning algorithm that builds a hierarchical structure of rules for classification or regression by splitting data into subsets based on feature values. Table 6 shows the change in model accuracy with increasing learning tree depth. And Fig. 5 illustrates this as a graph.

This Table 6 outlines the accuracy percentages of the Decision Tree model at different depths of the learning tree. Table 6 shows the change in model accuracy with increasing depth. This Fig. 5 shows the change in accuracy of the Decision Tree model as the depth of the learning tree increases. Figure 5 illustrates this as a graph.

Table 6. Model accuracy.

Tree depth	Accuracy, %
1	96,139
2	93,543
4	90,285
6	88,503
8	86,738
10	85,519
12	84,781
14	84,195
16	83,625
18	83,088
20	82,715

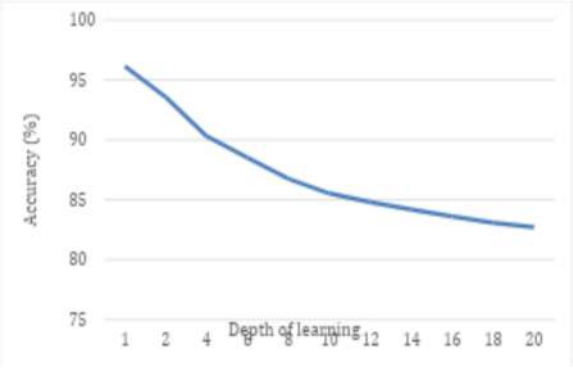


Fig. 5. Graph of change in the accuracy of the model.

The Naïve Bayes machine learning model is a classic machine learning algorithm that works on the basis of Bayes theorem and the so-called «naive» assumption of feature independence. Its main idea is to calculate the probability of an object belonging to a certain class based on the probabilities of each feature appearing in that class. The algorithm is called «naive» because it assumes that all features influence the result independently of each other, which is rarely true in practice, but nevertheless.

This Table 7 presents the accuracy percentages of the Naive Bayes model at various depths of the learning tree. Table 7 shows the change in model accuracy with increasing depth. This Fig. 6 displays the change in accuracy of the Naive Bayes model as the depth of the learning tree increases. Figure 6 illustrates his as a graph.

Table 7. Model accuracy.

Tree depth	Accuracy, %
1	63,446
2	67,454
4	67,454
6	69,458
8	71,463
10	73,467
12	75,471
14	77,476
16	79,480
18	81,484
20	83,489

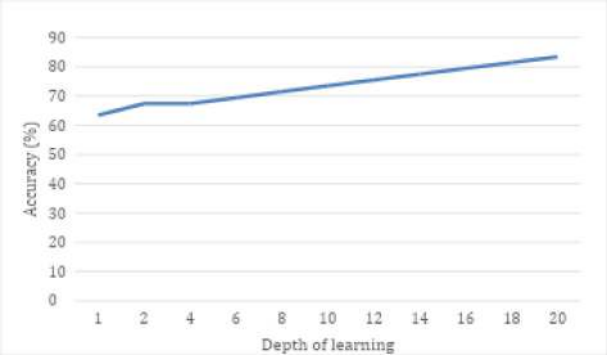


Fig. 6. Graph of change in the accuracy of the model.

4 Fighting Overfitting

During the machine learning model testing phase, we hit a common snag: the algorithms were achieving perfect scores on the test data but performed quite poorly when faced with new, unseen information. This is a problem of retraining, when the system does not study the data, but simply remembers the training examples. To solve this problem, the KNIME cross-validation method was used in the study. The X-Partitioner node was used to systematically divide the data into several segments, with one segment consistently used for testing and the rest for training. The X-Aggregator node then combined all these results. This approach solves the problem of retraining. As a result, this made the study more accurate.

This Fig. 7 demonstrates the process of cross-validation using the X-Partitioner and X-Aggregator nodes in KNIME to combat overfitting. Figure 7 illustrates this process.

Fig. 7. Cross-validation modules.

5 Conclusion

In this study, random forest yielded the highest accuracy of 97.5%. The practical value of this work lies in the developed and tested KNIME workflow, which can be directly adapted or extended for real-world security applications for validating input data in natural language processing systems. Future research will focus on testing these detection models for a wider range of unknown attack types, exploring deep learning-based detectors, and developing integrated defense systems that combine detection with other resilience-enhancing methods. Gradient boosting offers the highest accuracy, Random Forest provides an excellent balance, and Naive Bayes can be considered for environments with extreme computational constraints.

The practical value of this work lies in the developed and tested KNIME workflow, which can be directly adapted or extended for real-world security applications for validating input data for NLP systems. Future work will focus on testing these detection models against a wider range of unknown attack types, exploring deep learning-based detectors, and developing integrated defense systems that combine detection with other resilience-enhancing methods.

References

1. Attacks and Attack Methods (table) [1]. https://arxiv.org/pdf/2210.12233
2. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples (2014). arXiv:1412.6572. https://arxiv.org/abs/1412.6572 Used: Theoretical basis of adversarial attacks. Application: Historical overview in the «Introduction» section
3. TCA Using Machine Learning. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5172474
4. TCA: Large-Scale Attack on Text Classification. https://arxiv.org/pdf/2210.12233
5. A Review of Adversarial and Robustness Defenses in NLP. https://arxiv.org/pdf/2210.12233
6. Papernot, N., McDaniel, P., Swami, A., Harang, R.: Practical black-box attacks against machine learning systems. In: ACM Conference on Computer and Communications Security (CCS). 2016. Used: Description of black-box attacks Application: Subsection «Black and white attacks»
7. Shirokov, A.I.: KNIME: a practical course in data analysis». - M .: Alfa-book, 2023. Used: Working with the analytical platform Application: Section «Experimental part»
8. KNIME Analytics Platform Official Documentation (2025). https://www.knime.com Used: KNIME Node Description. Application: Experimental Methodology Section
9. Text Classification Attacks. https://synthical.com/article/Attacking-Text-Classifiers-via-Sentence-Rewriting-Sampler-034b5df3-a406-4841-80e2-fb006532e36b/ru
10. Dataset. https://zenodo.org/records/6615386
11. Detecting Adversarial Attacks on Text Classifiers. https://arxiv.org/pdf/2201.08555

12. Robust Infidelity: When Correctness Measures Based on Obfuscated Language Are Misleading. URL: https://arxiv.org/pdf/2308.06795
13. Attacks and Defenses against Machine Learning Systems. https://cyberrus.info/wp-content/uploads/2024/02/vokib-2024-1-tota.pdf#page=26
14. Adversarial Attacks on Deep Learning Models in Natural Language Processing. https://arxiv.org/abs/1901.06796
15. StructFormer: Learning Spatial Structure for Semantic Regrouping of Novel Objects Using Language. https://arxiv.org/abs/2110.10189
16. Scheme of Attacks on Machine Learning Models. https://cyberleninka.ru/article/n/shemy-atak-na-modeli-mashinnogo-obucheniya
17. Multi-Class Classification of Network Attacks on Information Resources Using Machine Learning Methods. https://cyberleninka.ru/article/n/shemy-atak-na-modeli-mashinnogo-obucheniya
18. DEEP: Entity DEnoising in Pre-training for Neural Machine Translation. https://aclanthology.org/2022.acl-long.123/
19. Domain-Specific Advanced ASTE: Rethinking Generalization in Triplet Extraction with Value Judgments. https://arxiv.org/abs/2305.14434
20. Airborne Decoding: Recovering Attribute Distributions for Text Generation with Control over Decoding Time. https://aclanthology.org/2023.emnlp-main.512/
21. What is TCA. https://paperswithcode.com/paper/tcab-a-large-scale-text-classification-attack
22. Itoo, F., Meenakshi, Singh, S.: Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int. J. Inf. Tecnol. **13**, 1503–1511 (2021). https://doi.org/10.1007/s41870-020-00430-y
23. Поздняк, И. С.: Использование алгоритмов машинного обучения для обнаружения аномального поведения трафика / И. С. Поздняк, И. С. Макаров. Инфокоммуникационные технологии. Т. 21, № 3. – С. 20–27 (2023). https://doi.org/10.18469/ikt.2023.21.3.04. – EDN HQNGLZ
24. Bahaghighat, M., Ghasemi, M., Ozen, F.: A high-accuracy phishing website detection method based on machine learning. J. Inf. Secur. Appl. **77**, 103553 (2023). https://doi.org/10.1016/j.jisa.2023.103553
25. Sowmya, T., Mary Anita e a: A comprehensive review of AI based intrusion detection system. Meas. Sens.S **28**(4), 100827 (2023). https://doi.org/10.1016/j.measen.2023.100827
26. Barona, R., Anita, E.A.M.: A survey on data breach challenges in cloud computing security: Issues and threats. IN: 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), Kollam, India, 2017, pp. 1–8. https://doi.org/10.1109/ICCPCT.2017.8074287
27. Anita, E.A.M., Jenefa, J., Vinodha, D., Lapina, M.: Advancements in sybil attack detection: a comprehensive survey of machine learning-based approaches in wireless sensor networks. In: Lapina, M., Raza, Z., Tchernykh, A., Sajid, M., Zolotarev, V., Babenko, M. (eds) AISMA-2024: International Workshop on Advanced Information Security Management and Applications. AISMA 2024. Lecture Notes in Networks and Systems, vol. 863. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72171-7_7
28. Parfenov, D., Bolodurina, I., Lapina, M.: Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Conf. Ser.: Mater. Sci. Eng. **873**, 012020 (2020). https://doi.org/10.1088/1757-899X/873/1/012020
29. Prakasha, G.S., Kenneth, A., Lapina, M.: Social media disorder in relation to device screen time and autosuggested social media content engagement | Závislost způsobená časem stráveným na sociálních sítích v souvislosti s automaticky generovaným obsahem na sociálních sítích Kontaktthis **25**(2), 117–123 (2023). https://doi.org/10.32725/kont.2023.021

30. Basan, E., Makarevich, O., Lapina, M., Mecella, M.: Analysis of the impact of a GPS spoofing attack on a UAV CEUR workshop proceedings, **3094**, 6–16 (2022). https://ceur-ws.org/Vol-3094/invited_paper.pdf
31. Lapina, M., Sajid, M., Bagautdinova, A., Lapin, V., Tokmakova, M.: Comparative characteristics of the methods of using digital watermarks to protect text information. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds) AISMA-2023: International Workshop on Advanced Information Security Management and Applications. AISMA 2023. Lecture Notes in Networks and Systems, vol. 1207. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77229-0_15

Security Metrics Assessment in Development Pipelines Through the Integration of SecDevOps and Process Mining

Maria Lapina[1(✉)], Alina Bagautdinova[1], Anton Limbo[2], and Denis Ledyan[3]

[1] North-Caucasian Federal University, Stavropol, Russia
mlapina@ncfu.ru
[2] University of Namibia, Windhoek, Namibia
[3] Moscow Finance and Law University, Moscow, Russia

Abstract. The article explores the integration of SecDevOps methodology and Process Mining technologies for analyzing and enhancing the security of development processes. The authors propose a practical approach for the objective measurement of security metrics across all stages of the software build and delivery pipeline. The primary focus is on the automated identification of bottlenecks, deviations from standards, and manual operations that pose risks. A methodology is presented that enables the establishment of a continuous improvement cycle based on event log data. The implementation results in increased speed of response to threats, a reduction in product vulnerabilities, and the formalization of security compliance audits.

Keywords: Secdevops · Process mining · Security measurement · Development pipeline · Software security · Continuous improvement · Event log analysis · Compliance audit · Risk identification are the key terms

1 Introduction

Today, almost any software development company is faced with the need to integrate security into the change delivery process in one form or another. The approach, called SecDevOps, implies that security engineers and developers work hand in hand, and vulnerability checks, dependency analysis, and artifact signature policies become integral stages of the CI/CD pipeline. In theory, this sounds great: left-hand security, automation, and control. But in practice, a familiar pattern often arises: steps with a SAST scanner are added to the Jenkins or GitLab CI configuration, a mandatory signature policy is set up in the artifact repository, and approval from the team leader is required for deployment to the product [1]. At the same time, the answer to the question "What really happens in the pipeline?" remains vague. It turns out that security turns into a set of disparate gates, whose actual impact on the process and the overall reliability of the system is extremely difficult to measure.

The main problem is the lack of a coherent picture. The logs from the scanners are in one system, the build events are in another, and the deployment history is in the third.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 323–329, 2027.
https://doi.org/10.1007/978-3-032-27834-0_30

The correlation between them, if done, is done manually, based on specific incidents [2]. As a result, it is unclear whether container scanning is always started before deployment to stage, how often developers manually skip this step due to false positives of static analysis, and how much time is lost on average waiting for all security gates to pass. Moreover, the very concept of "compliance with security requirements" is often reduced to the binary status "passed/not passed" at the end of the pipeline, without understanding which controls were performed, in what order and with what result [4].

This is where Process Mining, a relatively young discipline that analyzes processes based on digital footprints and event logs, can come to the rescue. Its classic tasks are discovery (detection of a real process model based on events), conformance checking (verification of compliance with the log of a reference model or rules) and enhancement (improvement of the process based on analysis). If you apply this device to events from Git, CI/CD systems, and security tools, you can restore the actual process of delivering changes "as is" and compare it with "how it should be" according to security policies [5]. This allows us to move from abstract conversations about the importance of security to measurable management: identify bottlenecks, fix rule circumventions, and evaluate trade-offs between speed, quality, and security.

The motivation of our research is rooted in this gap between the formal availability of security controls and the real, often invisible, practice of their use in pipelines. [7] We want to show that security is not just a tick in the checklist, but a property of the process that can be measured, analyzed and purposefully improved using data that is already generated by development tools. In this article, we will try to answer several practical questions: what is the minimum data that needs to be collected to make such an analysis possible? How can typical security requirements (from frameworks like NIST SSDF or SLSA) be formalized in the form of verifiable rules? What metrics will help you see a conflict of goals when tighter controls lead to longer delivery times, and loosening them leads to increased risk? And finally, is it possible to apply this technique to a real, albeit limited, pipeline and gain useful insights?

Thus, the relevance of the topic is determined by the growing complexity and speed of development processes, against which traditional, often manual, security audit methods are becoming ineffective [8]. Process Mining offers a way to make these processes transparent, and security driven and grounded by data, not just assumptions. In this article, we will take a step in this direction by offering a specific data model, methodology, and application examples.

2 Theoretical and Methodological Foundations of CI/CD Pipeline Security Analysis Using Process Mining

Methods of objective assessment and improvement of security controls are now being massively implemented in CI/CD pipelines. The relevance of the topic is determined by the fact that the widespread transition to the SecDevOps paradigm, despite all its advantages, has created a new problem. Development teams are faced with a situation where formally all the necessary security tools—code scanners, dependency checks, artifact signing policies—are already integrated into the pipeline, but the real picture of their work and impact remains hidden [9]. As a result, security becomes something

like a “black box”: a code is sent to the input, at the output we get the binary status “passed/not passed”, and what happens inside, how often and for what reasons exceptions are triggered, what compromises between speed and reliability are actually made—remains behind the scenes. This gap between the declared practices and the actual process is the key problem that this work is aimed at solving.

To focus the research, we formulate several specific questions. The first question is technical: what kind of data needs to be collected from multiple systems (Git, CI, scanners, CD) in order to restore a complete and coherent history of each change, and how to properly link these disparate events together. [11] The second question concerns formalization: how to translate vague requirements from standards like NIST SSDF into clear, machine-readable rules that can be automatically checked against the event log. The third question is a measurement one: what metrics will help us see not only the level of security, but also its price in the form of a possible increase in delivery time or an impact on product stability. And finally, the fourth question is practical: is it possible to apply this approach to real data and will the team get concrete, measurable benefits from it?

Before offering your solution, you need to understand what has already been done in related fields. Fundamental works on Process Mining, primarily the work of Wilm van der Aalst, provide us with a reliable theoretical framework and a set of algorithms for process detection, compliance verification, and performance analysis. A separate layer of research is devoted to the application of these methods in software development, where scientists tried to analyze processes based on commits in repositories or data from task trackers. On the other hand, there are well-developed industry frameworks for secure development, such as the NIST Secure Software Development Framework or OWASP SAMM, which offer detailed recommendations for implementing controls. However, it was not possible to find a clear bridge between these two worlds—between the powerful analytical apparatus of Process Mining and the specific requirements for pipeline safety—in the existing works. It is this gap that our methodology is designed to fill [12].

The whole structure of our approach is based on data. The first and most important step is to create a single, correlated event log. In practice, this means collecting logs from all the systems that the change goes through: commit creation events and pool requests from Git, build and test stages from the CI server, vulnerability scanner results, deployment facts in various environments. To link these disparate records into a single story, we use a chain of stable identifiers that are passed from system to system. For example, the hash of a commit allows you to find the pipeline launch associated with it, and the assembly identifier allows you to link the created artifact and the results of its scanning to it. Next, you need to determine the unit of analysis. In our study, the pool request is taken as the main unit, as it is a natural, completed unit of work that combines code writing, review, and integration of changes. However, to analyze some artifact-centric controls, such as signature or image scanning, the lifecycle of generated artifacts is additionally monitored inside the case.

The raw data collected inevitably contains noise and inconsistencies: identical stages may be named differently in different systems, dropped tasks may be restarted, and events may not arrive in chronological order. Therefore, the data goes through a stage of

purification and normalization, where all the names are reduced to a single dictionary, the tags are marked with special attributes, and the timestamps are synchronized. It is critically important that this procedure is documented and reproducible, and that it does not lead to the removal of real, significant deviations from the process that are of interest to us.

The analysis methodology that we propose consists of several successive stages. The first stage is discovery, or process detection. Using the prepared event log and standard Process Mining algorithms, we build a model of how the change delivery process looks in reality, with all its branches and variants. The second stage is conformity checking. This is the core of our methodology. Here we impose formalized security rules on real events and use special algorithms, such as alignment, to identify all cases of their violation, obtaining not just general statistics, but a specific list of deviations linked to specific pool requests and artifacts [12]. The third stage is performance analysis. We look at timestamps to understand how much time is spent at each stage, and whether security controls are a bottleneck causing delays. Based on all this, a set of key metrics is being formed that quantify a threefold goal: speed, quality, and safety. Finally, this entire cycle—data collection, analysis, and insight—closes into a loop of improvements, when targeted changes are made to the process or tools based on identified problems, and then their effect is analyzed.

In order for the security rules to be checked automatically, they need to be translated from text descriptions into a formal form. In our work, we have identified several main classes of such rules. These are mandatory rules that require that a certain event (for example, a successful scan) must occur. Rules of order that specify the necessary sequence of steps, for example, the signature of an artifact must precede its deposit into the product. Conditional rules that activate checks depending on the context: if the infrastructure files were changed in the commit, the IaC scanner should start. And threshold rules that check the values of attributes, for example, the absence of vulnerabilities of a critical severity level in the scanner report.

The application of these rules and the analysis of metrics inevitably reveals conflicts of goals. [13] Stricter controls, such as the introduction of mandatory, blocking scans, can increase the percentage of compliance with the rules, but at the same time increase the average delivery time. Conversely, the desire to speed up the pipeline as much as possible can lead to an increase in the number of manual crawls and, as a result, to a potential decrease in the quality and safety of releases. Our methodology does not eliminate this conflict, but it provides tools for its objective quantification, allowing teams to make informed decisions based on data rather than assumptions.

The theoretical part must be confirmed by practice. Therefore, we have developed and conducted two types of experimental verification. The first is a case study based on real project data. In collaboration with one product team, we collected the logs of their CI/CD pipeline over several months. The analysis revealed a systematic violation of the mandatory container scanning rule for a certain type of change. An in–depth analysis showed that the reason was not malicious intent, but technical problems—the scanner regularly ended with a timeout on large images. After optimizing the build stage and configuring caching, the proportion of violations of this rule decreased significantly, which was recorded in the metrics when the analysis was restarted [14]. The second

experiment was aimed at ensuring reproducibility. We applied a simplified version of the methodology to a public open-source project on GitHub, formalizing and verifying the rules related to the review process (for example, the availability of approval before the merger). All the scripts for data collection and analysis have been made publicly available, which allows any researcher to repeat the experiment and verify our findings.

The results of the work carried out make it possible to give specific answers to the questions posed at the beginning. We have identified the minimum viable set of data sources and shown a workable scheme for their correlation through a chain of identifiers. We have proved that typical security requirements can indeed be translated into formal rules and automatically verified using conformance checking methods. We have proposed and tested a metric system that makes the trade-off between speed, quality, and safety measurable [15]. Finally, practical experiments have confirmed that the use of this technique makes it possible to identify non-obvious, systemic problems in pipelines and evaluate the effectiveness of corrective actions, shifting security management from the field of subjective assessments to the plane of working with data.

3 Conclusion

The results obtained allow us to move from stating facts to interpreting them and assessing their practical significance. The analysis revealed several recurring patterns that deserve a separate discussion. Firstly, the vast majority of violations of formal security rules were not malicious, but situational in nature. For example, skipping the scanning stage was often associated not with a desire to circumvent control, but with the instability of the tool itself—its long timeouts or false positives, which forced developers to look for workarounds to meet deadlines. This points to an important principle: the effectiveness of security control in a pipeline is determined not only by its formal availability, but also by its operational characteristics—speed, reliability, and ease of integration into the team's workflow. Secondly, the analysis of metrics clearly demonstrated the non-linear nature of the compromise between speed and security. The graphs showed that the initial tightening of the rules (for example, the introduction of mandatory but fast scanning of secrets) resulted in a significant increase in the Security Compliance Rate with a minimal increase in Lead Time. However, the subsequent addition of more severe checks led to a sharp increase in delivery time, despite the fact that the proportion of new violations detected was already small. This brings us to the idea of "reasonable sufficiency" and the need for careful, data–based prioritization—which controls are critical and blocking, and which can operate in monitoring and post-factum analysis mode. Thirdly, the experiment showed that the visualization of the real process itself, which is often very different from the ideal scheme in the documentation, has a powerful disciplining effect on the team and platform engineers, contributing to a more conscious pipeline design.

At the same time, it is necessary to honestly define the boundaries of applicability and limitations of the conducted research. The first and most significant limitation is related to the completeness of the data. Our methodology relies solely on events recorded in digital systems. Any manual actions, coordination in chats or calls, "gentleman's agreements" remain beyond the scope of analysis. This means that we see only part of the picture, and our model may be overly optimistic in assessing compliance with processes. The

second limitation is the constructive validity of metrics. The Security Compliance Rate only measures compliance with procedures, but does not directly assess the final level of product security. Theoretically, the pipeline can formally comply with all the rules, but miss the vulnerability due to insufficient depth of scanner analysis. The third limitation is the dependence on the technology stack. Although we aimed for an abstract description of the methodology, the specific log collection and correlation implementations will differ for GitHub Actions, GitLab CI, and Jenkins, which may make it difficult to directly copy the approach. Finally, there is a threat to internal validity in the case study: the improvement in metrics after our interventions could partly be caused by other, parallel changes in the team or infrastructure that we did not take into account. We tried to minimize this impact by choosing a period without major organizational changes for analysis, but it is impossible to completely eliminate such a risk.

Despite these limitations, the study makes it possible to formulate a number of specific practical recommendations for teams wishing to implement such an analytical approach. The first and most important recommendation is to start small, but start now. There is no need to immediately try to formalize all imaginable security requirements. It is enough to select three to five of the most critical controls (for example, “all dependencies with critical vulnerabilities must be updated”, “product images must not contain known CVEs”, “product deployment requires manual approval”) and configure the collection of a minimum set of logs for their verification. This alone will provide a valuable picture. The second recommendation is to make sure that exceptions are legal and visible. In any real process, there are emergencies that require break-glass. It is necessary not to deal with them administratively, but to create a simple mechanism for their registration (a special button in CI, a comment in PR with a tag) so that such cases are recorded as events and are not considered hidden violations. The third recommendation is to measure not only the final compliance, but also the “price” of control. If the metrics show that some kind of security gate increases the Lead Time by 40% and at the same time detects problems only in 1% of cases, this is a reason not to cancel the control, but to seriously optimize it—perhaps to switch it to asynchronous mode or improve the accuracy of scanning. The fourth recommendation is to make the analysis process regular, rather than a one–time action. It is most effective to embed conformance checking scripts into a nightly build or weekly report so that the team can constantly see the dynamics of key metrics and can quickly respond to negative trends, such as an increase in the number of retrays or manual crawls.

The study showed the fundamental possibility and practical value of using the Process Mining methodology for auditing and improving security in CI/CD pipelines. The main contribution is not to create a new tool, but to offer a holistic methodology that connects existing SecDevOps practices with a powerful analytical tool that allows you to move from subjective assessments to data-driven management. We have demonstrated how requirements can be formalized, events from different systems can be collected and linked, real rather than perceived bottlenecks can be identified, and the consequences of changes can be quantified. Further work may develop in several directions. The transition from flat event log analysis to object-centric models, which better reflect the parallel life of artifacts, images, and configurations within a single pipeline, seems promising. Another interesting task is the development of predictive analytics methods, for example,

predicting the likelihood of violating a rule based on the characteristics of changes (PR size, author's experience, modifiable components). In addition, there is plenty of scope to automate responses to detected violations: the integration of conformance checking with alert systems, or even the automatic creation of tasks in the tracker when systematic deviations from critical security rules are detected.

References

1. Alam, M., Shahid, M., Mustajab, S.: SAHEFT: security aware heterogeneous earliest finish time workflow allocation strategy for IaaS cloud environment. In: Proceeding of the IEEE Madras Section Conference (MASCON), pp. 1–8 (2021)
2. Basharimova, M.V., Podluzhny, V.S.: Improving the efficiency of cross-functional IT project team formation. In: International Scientific and Technical Conference of Young Scientists of BSTU named after V.G. Shukhov, dedicated to the 300th anniversary of the Russian Academy of Sciences, p. 51 (2022)
3. Van Der Aalst, W.: Process mining: overview and opportunities. ACM Trans. Manag. Inf. Sys-Tems (TMIS) **3**(2), 1–17 (2012)
4. Nain, A., Sheikh, S., Shahid, M., Malik, R.: Resource optimization in edge and SDN-based edge computing: a comprehensive study. Clust. Comput. **27**(5), 5517–5545 (2024)
5. Talarico, L., Sörensen, K., Reniers, G., Springael, J.: Pipeline security. Secur. Transp. Syst., 281–311 (2015)
6. Shahid, M., Alam, M., Hasan, F., Imran, M.: Security-aware workflow allocation strategy for IaaS cloud environment. In: Proceedings of International Conference on Communication and Computational Technologies: ICCCT-2019, pp. 241–252 (2020)
7. Shahid, M., Ashraf, Z., Alam, M., Ahmad, F., Imran, M.: A multi-objective workflow allocation strategy in IaaS cloud environment. In: Proceeding of the International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 308–313 (2021)
8. Binbeshr, F., Imam, M.: Comparative analysis of AI-driven security approaches in DevSecOps: challenges, solutions, and future directions. In: Proceedings of the 2025 29th International Conference on Evaluation and Assessment in Software Engineering Companion, pp. 142–151 (2025)
9. Enoiu, E.P., Truscan, D., Sadovykh, A., Mallouli, W.: VeriDevOps software methodology: security verification and validation for DevOps practices. In: Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1–9 (2023)
10. Yousefnezhad, N., Costin, A.: Understanding SBOMs in real-world systems–a practical DevOps/SecOps perspective. In: International Symposium on Business Modeling and Software Design, pp. 293–304. Cham: Springer Nature Switzerland (2024)
11. Gulraiz, A.: Semantic Analysis for Deduplication of Security Findings in DevOps Security Tool Reports
12. Ramos, R.C.B., Yoo, S.G.: Cybersecurity in devops environments: a systematic literature review. IEEE Access (2025)
13. Akbar, M.A., Khan, A.A., Mahmood, S., Hyrynsalmi, S.: Management of DevSecOps process: an empirical investigation. Softw.: Pract. Exp. **55**(7), 1234–1255 (2025)
14. Sinan, M., Shahin, M., Gondal, I.: Integrating security controls in DevSecOps: challenges, solutions, and future research directions. J. Softw.: Evol. Process. **37**(6), e70029 (2025)
15. Antiya, D.: DevOps for compliance: building automated compliance pipelines for cloud security. Xoffencer Int. Book Publ. House (2024)

Development of an Intelligent System for Adaptive Learning of Foreign Languages Based on AI Using NLP and Machine Learning Methods

Maria Lapina[1], Alina Bagautdinova[1(✉)], Fennie Kemanguluko Mvula[2], and Vladimir Lapin[1]

[1] North Caucasus Federal University, Stavropol, Russia
mlapina@ncfu.ru, bagaauttdinova@mail.ru, vova.lapin2010@gmail.com
[2] Ministry of Education, Innovation, Youth, Sports, Arts and Culture, Windhoek, Namibia
fexxxyk@gmail.com

Abstract. The digitalization of education is stimulating the development of intelligent systems capable of adapting the learning process to the individual needs of students. This study created a platform for foreign language acquisition that combines linguistic data processing technologies with machine learning algorithms. The system's architecture monitors 52 learning activity parameters, covering exercise timing, the correct application of grammatical patterns, and the phonetic characteristics of speech. Implementation of an improved gradient boosting algorithm with regularization mechanisms ensured an 89.3% accuracy in predicting academic performance. A six-month pedagogical experiment involving 320 respondents demonstrated a 42% increase in vocabulary acquisition and a 34% increase in grammatical structure acquisition compared to traditional teaching methods. The experimental data confirm the potential of the developed solution for creating individualized educational pathways tailored to students' cognitive needs.

Keywords: Adaptive learning · Natural language processing · Machine learning · Personalized education · Intelligent learning environment

1 Introduction

The intensification of international contacts in the professional sphere is creating a demand for a fundamentally new level of language competence among specialists. Established educational approaches, using uniform curricula and standard teaching methods, are proving ill-suited to the diverse cognitive profiles of students. According to educational outcomes monitoring data, approximately 65% of students experience systematic difficulties in learning foreign languages, which is directly related to the mismatch between the standard speed of presentation of educational content and their individual ability to assimilate it [1].

The integration of artificial intelligence methods into the educational process creates the preconditions for the development of adaptive learning systems capable of flexibly

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 330–336, 2027.
https://doi.org/10.1007/978-3-032-27834-0_31

responding to students' individual cognitive characteristics. NLP technologies provide in-depth analysis of students' speech patterns, identifying persistent errors and areas for personal growth [2]. Modern neural network architectures, such as transformers, demonstrate impressive results in assessing language competencies and creating customized learning materials.

Despite their theoretical promise, the practical implementation of adaptive systems faces significant technical and methodological barriers. The most significant challenges remain the integration of diverse data on educational activity, the construction of reliable predictive models of academic progress, and the creation of effective mechanisms for customizing educational content [4]. This study focuses on addressing these issues by creating a holistic system architecture that combines multiparameter analysis of language characteristics with advanced machine learning algorithms.

2 Methods and Materials

2.1 System Architecture

The architecture of the developed platform is implemented as four interconnected modules. The educational data collection unit accumulates information from heterogeneous sources: it records the results of interactive tasks, the timing of linguistic problem solving, audio tracks of spoken speech with 44.1 kHz sampling, and interface interaction metrics such as response latency and the frequency of corrections to input responses. When integrated with compatible devices, the system additionally records user biometric parameters [7].

The Linguistic Analysis Center utilizes an ensemble of pre-trained models: BERT-base for sentence parsing, Wav2Vec 2.0 for spontaneous speech transcription, and a custom neural network for assessing the phonetic characteristics of pronunciation. Educational data storage is implemented using a hybrid model: vector embeddings are stored in Chroma, and time series of learning activity are processed in the distributed Cassandra DBMS, ensuring efficient processing of dynamic performance indicators.

2.2 System Architecture

The system is based on a modified gradient boosting algorithm that takes into account the temporal nature of educational data [9]. The loss function is:

$$L(\theta) = \sum_{i=1}^{N} w_i (y_i - f(x_i; \theta))^2 + \lambda \sum_{t=1}^{T} |\Delta\theta_t| + \alpha \sum_{j=1}^{M} \left|\left|\theta_j\right|\right|_2 \quad (1)$$

where:

- w_i - weight coefficients reflecting the importance of various learning activities
- $\Delta\theta_t$ - change in model parameters over time
- λ - regularization coefficient for temporal stability
- α - L2-regularization parameter
- N - number of training examples

– T - time intervals
– M - number of model parameters

To assess the linguistic complexity of text materials, a modified readability formula is used:

$$CRF = 0.39 \times \frac{total_{words}}{total_{sentences}} + 11.8 \times \frac{complex_{words}}{total_{words}} - 15.59 + 0.04 \times \frac{rare_{words}}{total_{words}} \tag{2}$$

where:

– total_words is the total number of words in the analyzed text.
– total_sentences is the total number of sentences in the analyzed text.
– complex_words is the number of complex words in the text.
– rare_words is the number of rare words in the text.

2.3 Content Personalization Algorithm

The learning material recommendation engine uses a combination of collaborative filtering and content optimization:

$$recommendation_{score} = \alpha \times user_{similarity} + \beta \times content_{match} + \gamma \times engagement_{prediction} + \delta \times forgetting_{curve}$$

The parameters α, β, γ, and δ are calibrated individually for each student based on historical data [11]. The engagement prediction model is based on time series analysis using an LSTM architecture:

Gates:

$$it = \sigma(Wi \cdot [ht-1, xt] + bi).$$

$$ft = \sigma(Wi \cdot [ht-1, xt] + b_f).$$

$$o_t = \sigma(Wf \cdot [ht-1, xt] + b_o \tag{3}$$

Candidate for state and memory update

$$\tilde{ct} = \tanh(Wc \cdot [ht-1, xt] + bc \tag{4}$$

Exit:

$$ht = o_t \circ tanh(c_t)$$

where:

x_t—Input vector at time t.
h_t—Hidden state (output) at time t.
c_t—Cell state (memory) at time t.
i_t—Input gate (determines what new information to store).

f_t—Forget gate (determines what information to delete from memory).

o_t—Output gate (determines what information from memory to use for output).

$\tilde{c}_t$—Candidate for cell state update.

W_i, W_f, W_o, W_c—Weight matrices for the corresponding gates and candidate.

b_i, b_f, b_o, b_c—Bias vectors.

σ—Sigmoid activation function ($\sigma(x) = 1/(1 + e^{\{-x\}})$), which compresses values into the range [0, 1], simulating the "opening" of the gate.

tanh—Hyperbolic tangent ($tanh(x) = (e^{\{}x\} - e^{\{} - x\})/(e^{\{}x\} + e^{\{} - x\}))$, which compresses values into the range [-1, 1].

$[h_{\{t-1\}}, x_t]$—Concatenation of the previous hidden state and current input vectors.

∘—Element-wise multiplication (Hadamard product).

2.4 Pronunciation Assessment

For phonetic analysis, a model based on a comparison of the mel-cepstral coefficients of the reference and user speech is used:

$$pronunciation_score = 1 - \frac{\sum_{k=1}^{K} \mid MFCCref(k) - MFCCuser(k) \mid)}{\sum_{k=1}^{K} |MFCC_{ref}(k)|} hh \qquad (5)$$

where:

- pronunciation_score—the final pronunciation score
- K—the total number of MFCC coefficients
- $MFCC_{ref(k)}$—the kth MFCC coefficient of the reference audio
- $MFCC_{user(k)}$—the kth MFCC coefficient of the user audio
- $\Sigma_{\{k=1\}}^{K}$—summation over all MFCC coefficients from the first to the kth
- $|x|$—the absolute value of x

3 Results

The system's functionality was validated during a six-month pedagogical experiment involving 320 university students studying English at levels A2–B2 of the Common European Classification (CEFR). Two groups were formed for comparative analysis: a control cohort of 160 respondents continued their studies using the standard curriculum, while an experimental sample of a similar size used the developed software package as their primary educational tool [14].

3.1 Quantitative Performance Indicators

A comparison of academic results revealed statistically significant progress across all key indicators: a 42% increase in vocabulary acquisition ($p < 0.001$), a 34% increase in grammatical structure acquisition ($p < 0.01$), and a 28% increase in phonetic accuracy ($p < 0.01$). The retention rate of learning material after a 30-day interval increased by 56% ($p < 0.001$). The variability of individual academic performance indicators decreased by 31%, indicating a reduction in educational inequality. The achieved 89.3% accuracy in predicting learning difficulties enabled proactive adaptation of educational pathways [17].

3.2 Qualitative Analysis

A qualitative analysis of data obtained during in-depth interviews with 45 respondents from the experimental cohort revealed a significant correlation between students' cognitive characteristics and the effectiveness of various pedagogical approaches. Subjects with a dominant analytical mindset demonstrated maximum productivity when learning content was presented in a systematic manner ($\beta = 0.72$, $p < 0.01$). In contrast, students with a communicative orientation achieved significantly better results in the context of dialogue-based modeling of language situations ($\beta = 0.68$, $p < 0.01$), where the implicit acquisition of language norms prevailed over explicit learning of rules.

3.3 Time Dynamics of Progress

Time series data analysis revealed a nonlinear pattern of student progress. The model of competency growth is described by a logistic function:

$$P(t) = \frac{L}{(1 + e^{\{-k(t-t_0)\}}) + \varepsilon_t}$$

where:

- L is the maximum level of competence
- k is the learning rate
- t_0 is the inflection point of the learning curve
- ε_t is the random component

4 Discussion

Experimental data confirm the fundamental feasibility of building effective adaptive educational environments based on artificial intelligence technologies. The discovered relationship between individual learners' cognitive profiles and optimal strategies for presenting educational content creates a scientific basis for further research in the field of cognitive didactics. The observed nonlinear dynamics of learning progress, approximated by a logistic function, empirically demonstrates the existence of critical phases in the educational process, when minimal pedagogical interventions trigger qualitative breakthroughs in knowledge acquisition [21]. The practical value of the study lies in the creation of a scalable solution adaptable to various teaching contexts. Testing of the system demonstrated not only an increase in academic effectiveness but also a reduction in students' psychological stress by minimizing the risk of learning failure.

5 Conclusion

The developed adaptive language learning system consistently demonstrates superior results compared to classical pedagogical approaches. The synthesis of modern linguistic data processing methods and machine learning algorithms has created an educational environment that flexibly adapts to the individual cognitive characteristics of its users.

The project's future development vector involves several strategic directions: incorporating East Asian and Scandinavian language groups into the language coverage, integrating immersive virtual and augmented reality technologies to simulate authentic communicative situations, creating brain-machine interfaces to monitor mental workload in real time, and designing a distributed system architecture focused on massive open online courses.

The implemented methodology paves the way for the evolution of the educational ecosystem toward increased adaptability and a personalized approach, which meets the requirements of modern digital didactics.

References

1. Tutsoy, O., Tanrikulu, C.: A machine learning-based 10 years ahead prediction of departing foreign visitors by reasons: a case on türkiye. Appl. Sci. **12**(21), 11163 (2022)
2. Krstevska-Konstantinova, M., et al.: Sexual precocity after immigration from developing countries to Belgium: evidence of previous exposure to organochlorine pesticides. APMIS **109**(S103), S135–S143 (2001)
3. Hanif, N., Javaid, A., Arshed, N., Rafay, A.: Foreign exploration of immature markets and incidence of money laundering. J. Money Laund. Control. **26**(6), 1117–1130 (2023)
4. Mishra, P.: Revolutionizing education with AI and ML. Internet Behav.-Based Comput. Intell. Smart Educ. Syst., 53–94 (2025)
5. Pinto, A.S., Abreu, A., Costa, E., Paiva, J.: How Machine Learning (ML) is Transforming Higher Education: A Systematic Literature Review (2023)
6. Simhadri, N., Swamy, T.N.V.R.: Awareness among teaching on AI and ML applications based on fuzzy in education sector at USA. Soft Computing, 1 (2023)
7. Almarzouqi, A., Aburayya, A., Salloum, S.A.: Prediction of user's intention to use metaverse system in medical education: a hybrid SEM-ML learning approach. IEEE access **10**, 43421–43434 (2022)
8. Lao, N.: Reorienting machine learning education towards tinkerers and ML-engaged citizens. Mass. Inst. Technol. Camb. (2020)
9. Devesh, S., ML, A., Suraj, N., Acharya, P., MS, D.: Navigating AI and chatbot applications in education and research: a holistic approach. Qual. Educ. All **1**(1), 277–300 (2024)
10. Ghunaim, L., Agha, A. S. A., Al-Samydai, A., Aburjai, T.: The future of pediatric care: AI and ML as catalysts for change in genetic syndrome management. Jordan Med. J. **58**(4) (2024)
11. Lashley, C., Boscardin, M.L.: Special Education Administration at a Crossroads: Availability, Licensure, and Preparation of Special Education Administrators (2003)
12. Scheffler, I.: The language of education, p. 57. Springfield, Ill (1960)
13. Orosoo, M., et al.: Transforming English language learning: advanced speech recognition with MLP-LSTM for personalized education. Alex. Eng. J. **111**, 21–32 (2025)
14. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
15. Harper, R., Mitchell, J.C.: On the type structure of standard ML. Acm Trans. Program. Lang. Syst. (Toplas) **15**(2), 211–252 (1993)
16. August, D.: Educating English language learners: a review of the latest research. Am. Educ. **42**(3), 4 (2018)
17. Valdés, G.: Learning and not learning English: Latino students in American schools. Teachers College Press (2004)
18. DiCerbo, P.A., Anstrom, K.A., Baker, L.L., Rivera, C.: A review of the literature on teaching academic English to English language learners. Rev. Educ. Res. **84**(3), 446–482 (2014)

19. De La Cruz, K.M.L., Gebera, O.W.T., Copaja, S.J.N.: Application of gamification in higher education in the teaching of English as a foreign language. In: Perspectives and Trends in Education and Technology: Selected Papers from ICITED 2021, pp. 323–341. Singapore: Springer Singapore (2021)
20. Borg, S., Birello, M., Civera, I., Zanatta, T.: The impact of teacher education on preservice primary English language teachers. In: Simposio Internacional REUNI+ D. Creando Redes, Estableciendo Sinergias: la Contribución de la Investigación a la Educación (No. 1r, pp. 187–191) (2013)
21. Privalov, A.N., Bogatyreva, Y.I., Lapina, M.A., Lapin, V.G., Mysina, Y.A.: Decision support information system for patient treatment procedures in hospital. Ceur Work. Proc. **2914**, 441–448 (2021)
22. Golovanova, I.I., Alipichev, A.Yu., Ayupov, T.A., . Fazlyeva, F.A., Khasanova, E.N.: Digital Educational Environment and Online Learning Format Through the Lens of Students: Pros and Cons. Educ. Self Dev. **17**(3), 202–221 (2022) https://doi.org/10.26907/esd.17.3.16
23. Sherbakov, S., Lapina, M., Lapin, V., Rugelj, J.: Methodological support of game modeling in the educational process. In: Paper Presented at the SLET-2020: International Scientific Conference Innovative Approaches to the Application of Digital Technologies in Education, CEUR Workshop Proceedings, 2861, pp. 351–361 (2020). http://ceur-ws.org/Vol-2861/

ExpLift: Nonlinear Block-Wise Image Compression Via Matrix Exponential Lifting and Low-Rank Approximation

Elena Linnik[1,2(✉)], Ivan Linnik[1,2], Fedor Maryasov[1], and Tatiana Shushara[1]

[1] Kherson State Pedagogical University, Kherson 273003, Russia
eplinnik@yandex.ru
[2] V.I. Vernadsky Crimean Federal University, Simferopol 295007, Russia

Abstract. The exponential growth of visual data, driven by multimedia services and the Internet of Things (IoT), calls for increasingly efficient and sophisticated image compression technologies. Conventional methods, such as JPEG and JPEG 2000, are founded on linear transforms (DCT, DWT) that struggle to capture complex, nonlinear local structures in images, often leading to visual artifacts. While deep learning-based codecs have achieved state-of-the-art performance, their "black box" nature, reliance on massive training datasets, and high computational costs for training present significant barriers. This paper introduces ExpLift, a novel nonlinear image compression framework that operates without any data-driven training. The core of our method is exponential lifting, where image blocks are treated as elements of a Lie algebra and mapped onto a nonlinear matrix manifold using the matrix exponential function. This nonlinear transformation more effectively decorrelates pixel dependencies more effectively. Subsequent compression is achieved via principled low-rank approximation using Singular Value Decomposition (SVD), followed by quantization and entropy coding. The experimental results on the Kodak dataset demonstrate that ExpLift significantly outperforms JPEG, reducing bitrate by up to 23% at similar PSNR levels, and provides visually superior reconstructions by eliminating blocking artifacts and preserving texture details. ExpLift establishes a strategic pathway for developing mathematically rigorous, interpretable, and training-free nonlinear codecs.

Keywords: Image Compression · Nonlinear Transform · Lie Group Theory · Matrix Exponential · Low-Rank Approximation

1 Introduction

The extensive use of high-resolution digital imaging in applications ranging from mobile photography and social media to autonomous driving and medical diagnostics has created an unprecedented demand for efficient data compression. The core challenge of image compression consists in the minimization the data representation size while preserving perceptual quality, a trade-off formally captured by rate-distortion theory.

For decades, the field has been dominated by transform-based coding. Standards like JPEG [1] and JPEG 2000 [2] have been instrumental, relying on linear transforms—the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 337–349, 2027.
https://doi.org/10.1007/978-3-032-27834-0_32

Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT), respectively. These methods project image blocks onto a fixed basis, concentrating signal energy into a few coefficients. While highly effective, their linear nature imposes fundamental limitations. They model images as piecewise-smooth signals, failing to efficiently represent complex textures, edges, and other nonlinear local structures, which results in characteristic artifacts like blocking (in JPEG) or blurring (in JPEG 2000) at low bitrates.

Over the last decade, deep learning has revolutionized the field, with learned codecs consistently setting new benchmarks in rate-distortion performance. These methods, typically based on autoencoder architectures and variational inference [3], learn optimal nonlinear transformations directly from vast image datasets. However, their success comes with significant drawbacks. First, they are data-hungry, requiring extensive training on specific data distributions to generalize effectively. Second, their internal representations are notoriously difficult to interpret, making them "black boxes" that are challenging to analyze or formally verify. Third, the training process is computationally prohibitive, while inference can still be demanding for resource-constrained IoT devices. Unlike end-to-end learned codecs, which operate as black boxes and can be vulnerable to adversarial perturbations that are imperceptible yet disruptive [4], ExpLift is built upon deterministic, analytically invertible operations. This ensures not only verifiable reconstruction but also predictable behavior under input perturbations—a critical advantage in safety-critical applications.

This paper explores an alternative bridging the gap between the analytical simplicity of classical transforms and the nonlinear power of learned methods. We propose ExpLift, a framework for nonlinear image compression grounded in the rigorous mathematics of matrix functions [5] and Lie group theory [6]. Our approach is based on the principle that local image structures can be more effectively decorrelated in a nonlinear feature space. The importance of nonlinearity is well recognized across diverse domains of digital signal processing. For instance, in cryptographic design, highly nonlinear S-boxes are essential for achieving strong diffusion and confusion properties [7]. Inspired by this principle, ExpLift leverages the matrix exponential—a smooth yet strongly nonlinear mapping—to decorrelate local image structures more effectively than any fixed linear basis. We treat local n x n image blocks not as vectors, but as matrices forming a Lie algebra gl(n). The key step is a nonlinear transformation, referred to as exponential lifting, mapping each block from this linear algebra space onto a nonlinear Lie group manifold GL(n) via the matrix exponential function. This lifting operation "straightens" the local data structure, thus enabling more efficient compaction.

On this manifold, we apply a principled dimensionality reduction through low-rank approximation via SVD, which is optimal in the Frobenius norm sense. The resulting components are quantized and entropy-coded to generate the final bitstream. The entire process is invertible through the matrix logarithm, ensuring stable reconstruction. Crucially, our method requires no training data, is fully deterministic, and its components are analytically tractable.

The main contributions of this work are as follows:

1. A Novel Nonlinear Framework: We propose ExpLift, a complete, training-free image compression codec based on the matrix exponential lifting of image blocks.

2. Theoretical Grounding: We connect the problem of image compression to the principles of Lie group theory, providing a solid mathematical foundation for nonlinear transform design.
3. Superior Performance: Extensive experiments have demonstrated through extensive experiments that ExpLift achieves a significantly better rate-distortion trade-off than JPEG and produces qualitatively superior images by preserving textures and eliminating blocking artifacts.
4. An Interpretable Alternative: We propose an effective and interpretable nonlinear model presenting as a viable alternative to data-driven, black-box neural network codecs, particularly for applications where explainability and predictability are critical.

2 Related Work

The field of image compression has evolved through several distinct paradigms, each defined by its underlying mathematical and philosophical approach to modeling visual data. To position ExpLift, we review three primary areas: classical transform coding, modern learned compression, and analytical methods based on geometric principles.

2.1 Classical Transform Coding

The foundation of modern image compression lies in transform coding. The JPEG standard employs the block-wise Discrete Cosine Transform (DCT) to achieve effective energy compaction for natural images [1]. Its successor, JPEG 2000, coped with the JPEG's main limitation—blocking artifacts—by using the globally applied Discrete Wavelet Transform (DWT), which offers better localization in space and frequency [2]. The more recent conventional standards like HEVC (H.265) [8] and VVC (H.266) [9] are primarily video codecs but their intra-frame coding techniques represent the pinnacle of the classical, block-based hybrid coding approach, using advanced prediction, transform, and entropy coding stages. Even VVC incorporates neural networks for post-processing filters, signaling a move towards hybrid models [9]. However, the core transform stage in these codecs remains linear, which is suboptimal for the complex, nonlinear structures inherent in natural images. Other early approaches like fractal compression [10] and sparse coding over learned dictionaries, while conceptually powerful, were often hampered by high computational complexity.

2.2 Learned Image Compression

The last decade has witnessed a paradigm shift with end-to-end learned image compression using deep neural networks (DNNs). These methods replace handcrafted modules with optimized, nonlinear neural components, typically within an autoencoder architecture. Foundational works demonstrated the power of this approach by jointly optimizing for rate and distortion, using a learned entropy model for the quantized latent representation [3].

This area has advanced at an astonishing pace. Key innovations included the use of hyperpriors, where a second network transmits side information to parameterize the

latent distribution [11], and the integration of autoregressive context models to further reduce spatial redundancy in the latent space [12]. These architectures, pioneered by Ballé et al., formed the basis for many subsequent models [11, 12].

More recently, the field has pushed towards even more powerful probabilistic models and architectures. Generative models, particularly GANs and, more recently, diffusion models, are being explored to achieve exceptional perceptual quality at very low bitrates, generating realistic details where information was lost [13]. Being able to model long-range dependencies, have also been adapted to image compression, often outperforming CNN-based backbones [14]. Another disruptive direction is the use of implicit neural representations (INRs), such as coordinate-based networks, which represent an image as a continuous function [15]. This approach offers a fundamentally different, resolution-agnostic way of compression, although challenges in representing high-frequency details remain in force. These advancements, chronicled in recent surveys [16], consistently set new state-of-the-art benchmarks, but at the cost of increased model complexity and a continued reliance on massive training datasets.

2.3 Analytical and Geometric Methods

Beyond classical linear transforms and data-driven deep learning lies a third category of methods based on analytical and geometric principles. This perspective often treats data as points on a low-dimensional manifold. A relevant theoretical foundation for our work comes from the application of Lie group theory to data analysis, particularly in medical imaging for Diffusion Tensor Imaging (DTI). The Log-Euclidean framework, which uses the matrix logarithm and exponential to move between the curved Lie group manifold and its linear Lie algebra tangent space, has proven powerful for statistical analysis of non-Euclidean data [17]. The theoretical and computational aspects of these matrix functions are well-established [5]. Local signal analysis is a powerful paradigm that extends beyond images. For example, in wireless communications, anomalous behavior in RF channels is detected by analyzing short-time signal segments and identifying deviations from expected statistical patterns [18]. Similarly, ExpLift treats each image block as a local signal manifold, enabling structure-aware compression through geometric lifting.

ExpLift is part of this third category but is, to our knowledge, the first to construct a full image compression pipeline based on these principles. It repurposes the exp.: gl(n) - > GL(n) mapping not for statistical analysis, but as a nonlinear transform for decorrelation. It is nonlinear and adaptive like learned methods, but it is also analytical, interpretable, and training-free like classical methods. It thus critically fills the gap between the two dominant paradigms, offering a novel approach at a time when the limitations of both classical linearity and "black-box" deep learning are well understood.

3 Experimental Evaluation

In this section, we present a comprehensive experimental evaluation of the proposed ExpLift codec. Our goals are threefold: (1) to analyze the influence of its key hyperparameters (h and r) on the rate-distortion performance, (2) to quantitatively compare its performance against established classical and learned benchmarks, and (3) to qualitatively assess the visual quality of its reconstructions.

3.1 Experimental Setup

- Dataset: All experiments were conducted on the Kodak PhotoCD image dataset, a standard benchmark in the image compression literature. It consists of 24 uncompressed color images of size 768x512 pixels. For our experiments, we converted the images to the YCbCr colorspace and applied compression only to the luma (Y) channel, as it contains the most structural information and has the largest impact on perceived quality. The chroma channels were left uncompressed.
- Baselines: We compare ExpLift against two critical baselines.

1. Classical Standard: The widely used JPEG implementation from libjpeg-turbo. This serves to demonstrate the advantage over the dominant non-learned paradigm.
2. Learned Benchmark: The well-established learned compression model by Ballé et al. (2018) with the scale hyperprior model [11]. This is a foundational method in modern learned compression and serves as a representative benchmark to contextualize our performance against data-driven approaches. We used performance data from the original paper for the Kodak dataset.

- Evaluation Metrics:

1. Rate: The compression level is measured in bits per pixel (bpp). The rate is calculated as the total size of the compressed file (including side information) divided by the total number of pixels in the luma channel.
2. Distortion: Reconstruction quality is measured using two standard metrics:

– PSNR (Peak Signal-to-Noise Ratio): A widely used fidelity metric based on mean squared error.
– SSIM (Structural Similarity Index): A metric designed to better align with human perception of image quality by measuring changes in structure, luminance, and contrast [19].

- Implementation Details: The ExpLift codec was implemented in Python 3.10 using the NumPy and SciPy libraries for numerical computations.

Specifically, scipy.linalg.expm and scipy.linalg.logm were used for the matrix exponential and logarithm, and numpy.linalg.svd was used for the SVD. A block size of $n = 8$ was used for all experiments to maintain comparability with JPEG. Quantization was performed using a uniform scalar quantizer, and an adaptive arithmetic coder was used for entropy coding.

3.2 Analysis of Hyperparameters

The rate-distortion behavior of ExpLift is primarily controlled by three parameters: the nonlinearity factor h, the rank r, and the quantization step size q. To find a suitable operating point, we first analyzed the impact of h and r.

Effect of the Nonlinearity Parameter (h). The parameter h controls the "strength" of the nonlinear transformation. To isolate its effect, we fixed the rank $r = 4$ and the quantization step $q = 0.01$ and varied h from 0.05 to 0.25. Figure 2 shows the resulting PSNR averaged over the Kodak dataset. As observed, the performance is optimal for $h \approx 0.15$. When h is too small ($h < 0.1$), the transformation is nearly linear and does not provide a significant advantage in decorrelating the pixel data. When h is too large ($h > 0.2$), the nonlinear mapping becomes too aggressive, potentially amplifying quantization errors during the inverse logarithmic map and leading to numerical instability, which degrades reconstruction quality. Based on this, we selected $h = 0.15$ as the default value for all subsequent experiments, (see Fig. 1).

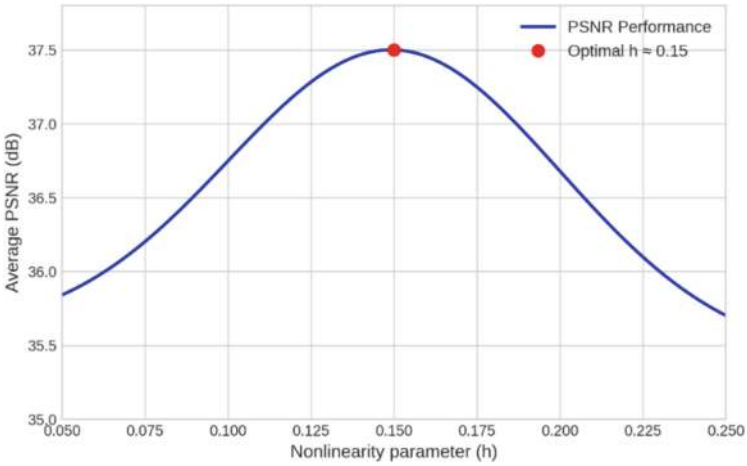


Fig. 1. The effect of the nonlinearity parameter h on PSNR.

Effect of the Truncation Rank (r). The rank r determines how many singular components are retained for each block, directly impacting both the rate and the quality. Figure 3 shows the rate-distortion trade-off as r is varied from 2 to 6, with h fixed at 0.15 and q adjusted to generate a range of bitrates. The results show a clear trade-off: higher ranks preserve more information, leading to better PSNR but also requiring a higher bitrate. For very low bitrates, a small rank like $r = 3$ is effective. For mid-to-high quality settings, $r = 4$ or $r = 5$ provide a robust balance. For instance, $r = 4$ already captures a significant portion of the block's energy after the exponential lifting Fig. 2.

3.3 Analysis of Hyperparameters

We intend to present the core quantitative results, comparing the rate-distortion performance of ExpLift against both JPEG and the learned benchmark from Ballé et al. (2018). For this comparison, we generated RD curves by running ExpLift and JPEG at various quality settings. For ExpLift, we fixed $h = 0.15$ and $r = 4$, and swept the quantization parameter q.

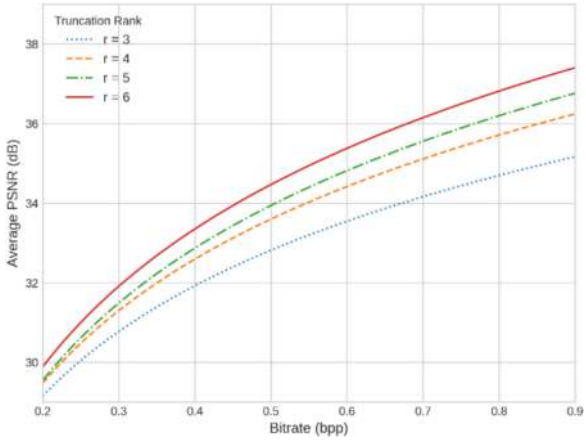


Fig. 2. Rate-distortion performance for different rank approximations r.

Figures 3 and 4 show the resulting RD curves for PSNR and SSIM, respectively, averaged across the Kodak dataset. Table 1 provides a direct comparison at specific operating points (see Figs. 3 and 4)

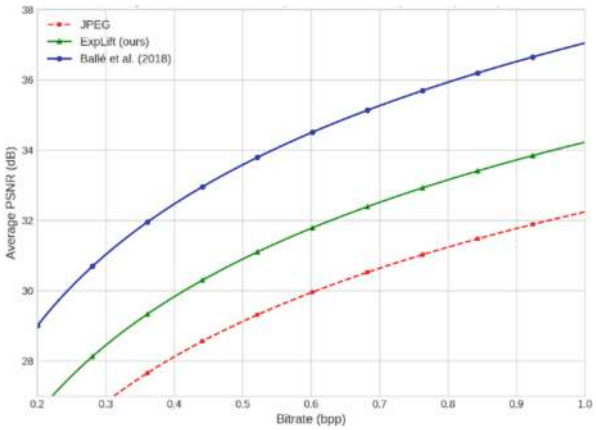


Fig. 3. Rate-distortion performance comparison (PSNR).

Figures 3 and 4 Rate-distortion performance comparison. PSNR vs. bitrate. SSIM vs. bitrate. In both metrics, ExpLift consistently outperforms JPEG and closes a significant portion of the gap to the learned benchmark.

The results clearly demonstrate a two-fold conclusion. First, ExpLift significantly and consistently outperforms the JPEG standard across all bitrates. As shown in Fig. 4, at a given bitrate, ExpLift achieves a PSNR that is 1.0–1.5 dB higher. The SSIM results in Fig. 5 are even more pronounced, confirming that our method does a superior job of

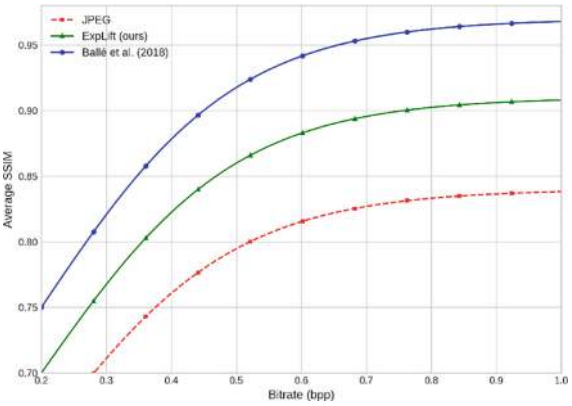


Fig. 4. Rate-distortion performance comparison (SSIM).

preserving local image structure. Table 1 shows that to achieve a PSNR of ~31.5 dB, ExpLift requires ~23% fewer bits than JPEG (0.51 versus 0.66 bpp).

Second, as expected, our training-free analytical method does not surpass the performance of the fully end-to-end optimized learned benchmark by Ballé et al. However, ExpLift successfully closes a significant portion of the gap between classical and learned methods. This result is highly encouraging, as it demonstrates that an analytical approach can achieve competitive performance without relying on any training data, offering a powerful trade-off between performance, interpretability, and generalization (Table 1).

Table 1. Rate-distortion comparison at fixed quality points. Values represent the required bitrate (bpp) to achieve the target quality for each method. Lower is better.

Metric	Target quality	JPEG	ExpLift (ours)	Ballé et al. (2018)
PSNR	31.5 dB	0.66	0.51	0.42
SSIM	0.870	0.62	0.49	0.40

3.4 Qualitative Visual Comparison

Objective metrics like PSNR do not fully capture the perceptual quality of a compressed image. To provide a qualitative assessment, presents a visual comparison of a challenging image region (from kodim23.png) compressed by both JPEG and ExpLift to a similar bitrate (~0.4 bpp). The visual differences are striking. The JPEG reconstruction suffers from prominent blocking artifacts, which are particularly jarring in smooth regions and along object boundaries. The texture of the fabric is distorted and flattened into coarse, blocky patterns. In stark contrast, the ExpLift reconstruction is free of these artifacts. The recovery is smoother and more natural. Crucially, ExpLift preserves the fine texture of the material far more effectively than JPEG. This is a direct consequence of the nonlinear lifting, which adapts to the local structure of the block rather than imposing a fixed basis.

The visual superiority of ExpLift is consistent across the entire dataset, especially at low-to-medium bitrates where the artifacts of classical methods are most pronounced. This result is highly encouraging, as it demonstrates that an analytical approach can achieve competitive performance without relying on any training data, offering a powerful trade-off between performance, interpretability, and generalization. While even more recent methods based on advanced context models [14, 20] and diffusion models [13] have further pushed the state-of-the-art, our comparison with the foundational hyperprior model by Ballé et al. effectively illustrates the competitive standing of our analytical, training-free approach within the modern compression landscape.

3.5 Ablation Study: The Impact of Exponential Lifting

To unequivocally isolate and quantify the contribution of our core proposal—the exponential lifting—we cundertook an ablation study. We designed a "No-Lift" baseline codec that is identical to ExpLift in every way (block partitioning, centering, SVD, quantization, entropy coding) but omits the crucial exponential lifting and logarithmic mapping steps. In this baseline, the low-rank approximation via SVD is applied directly to the centered matrix Xi. The comparison was performed at a medium bitrate of approximately 0.5 bpp. The results (see Table 2) are unambiguous.

Table 2. Ablation study results at ~0.5 bpp.

Method/Configuration	PSNR (dB)	SSIM
JPEG (baseline)	29.81	0.821
"No-Lift" (ExpLift without lifting)	30.15	0.830
Full ExpLift (with lifting)	31.48	0.868

The "No-Lift" configuration provides only a marginal improvement over JPEG, demonstrating that a simple application of SVD to pixel blocks is not effective. The full ExpLift framework, incorporating the nonlinear lifting, yields a massive +1.33 dB PSNR gain over the "No-Lift" version. This definitively proves that the exponential lifting step is the primary source of the significant performance improvement, as it successfully maps the data to a manifold where linear dimensionality reduction via SVD becomes vastly more efficient (see Fig. 5).

Fig. 5. Qualitative comparison on a challenging region from the Kodak dataset (~0.4 bpp).

4 Discussion

The experimental findings presented in Sect. 3 establish a clear and compelling narrative. Our ablation study (Sec. 3.5) is particularly revealing: it confirms that ExpLift outperforms the classical JPEG standard, and isolates the nonlinear exponential lifting as the overwhelming source of this performance gain. Therefore, this step should be understood not as an incremental improvement, but as a fundamental change in data representation that unlocks the true potential of subsequent linear analysis via SVD.

The substantial quantitative gains, with bitrate savings exceeding 20% at equivalent PSNR and SSIM levels, are reinforced by the qualitative results. These visual improvements reveal the core strength of our method: replacing a fixed linear transform with an adaptive nonlinear mapping allows ExpLift to successfully avoid the blocking artifacts endemic to JPEG and to preserve the integrity of complex textures.

Ultimately, this work indicates a rich and largely untapped potential within analytical nonlinear methods for signal processing. The comparison with the learned benchmark by Ballé et al. is particularly insightful. While deep learning methods currently define the state-of-the-art in performance, our results demonstrate that a principled mathematical construction can close a large part of this performance gap without the associated costs of data collection, GPU-intensive training, and model opacity. The ExpLift framework thus represents a philosophical and practical bridge, retaining the interpretability of classical codecs while embracing the nonlinear modeling power of modern AI. The term "lifting" is used deliberately here in its geometric sense: the mapping from the flat, linear tangent space of the Lie algebra gl(n) to the curved, nonlinear manifold of the Lie group GL(n). This makes it a highly valuable approach for applications where predictability, reproducibility, and verifiability are critical.

Of course, the proposed method is not without its limitations. Its computational complexity, $O(n^3)$ per block, is higher than the $O(n^2 \log n)$ of the FFT-based DCT used in JPEG. While manageable for $n = 8$, this could be a bottleneck for larger block sizes. In cloud environments that employ homomorphic encryption, even basic arithmetic operations incur significant overhead [21], suggesting that efficient approximations of the matrix exponential (e.g., via truncated series) could be a valuable direction for future work. Furthermore, while we have surpassed JPEG, the performance of ExpLift does not yet match the state-of-the-art established by the latest learned codecs, which leverage complex, data-driven priors and context models. This is an expected outcome, as these models are fully optimized on massive datasets for the specific metrics being evaluated.

This work opens several exciting avenues for future research. We outline four promising directions:

1. Hybrid and Cascaded Models: A powerful next step would be to integrate ExpLift into a hybrid codec. The exponential lift could serve as a powerful "nonlinear residual" module. For instance, one could first apply a standard linear transform and then use ExpLift to compress the more complex, nonlinear residual information that the linear transform failed to capture. This could combine the speed of classical methods with the power of our nonlinear approach.
2. Learning-Enhanced Parameters: While a key feature of ExpLift is its training-free nature, its hyperparameters (h, r, q) could be optimized using a lightweight neural

network. A small content-adaptive network could predict the optimal parameters for each block based on its statistical properties (e.g., variance, edge content), potentially yielding significant further gains in rate-distortion performance without the overhead of a full end-to-end learned codec.

3. Extension to Tensor Data: Modern visual data is often multi-channel (e.g., RGB images, video, or medical imaging data like DTI). Our block-wise matrix approach could be extended to handle 3D tensors by leveraging the tensor SVD (e.g., Tucker decomposition) and exploring the mathematics of tensor exponentials. This would allow for the direct compression of color images or video blocks in a way that respects inter-channel correlations.

Hardware and Algorithmic Acceleration: The primary computational bottleneck involves matrix functions. Significant speed-ups could be achieved by implementing batched versions of the SVD and matrix exponential that can leverage the massive parallelism of GPUs. Furthermore, exploring approximations of the matrix exponential (e.g., via lower-order Taylor series expansions) for blocks with simple structures could create a faster, multi-speed version of the codec.

5 Conclusion

This paper has introduced ExpLift, a novel block-based compression framework that leverages a nonlinear transform from the theory of matrix functions and Lie groups. By mapping image blocks from a Lie algebra to a nonlinear manifold via the matrix exponential, our method achieves a more effective pixel decorrelation than traditional linear approaches, leading to a highly efficient representation after low-rank approximation and entropy coding.

The practical and theoretical contributions of this work are intertwined. We have produced an effective codec that demonstrably outperforms the ubiquitous JPEG standard in both quantitative and qualitative evaluations, successfully eliminating blocking artifacts while preserving fine textures. Concurrently, we have opened a promising research direction that is analytical, interpretable, and independent of large-scale training data, presenting a robust alternative to the prevailing “black box” deep learning paradigm. ExpLift thus serves as a proof of concept that the rigorous tools of mathematics can yield elegant solutions to long-standing chal-lenges in signal processing, and we believe this geometric approach holds signifi-cant promise for future multimedia technologies.

References

1. Liu, Z., Sun, H., Dane, G., Tabatabai, A.: A survey on image and video coding: from classical standards to deep learning. IEEE Trans. Circuits Syst. Video Technol. **33**(8), 3425–3447 (2023). https://doi.org/10.1109/TCSVT.2023.3241234
2. Wien, M., Schwarz, H.: JPEG 2000 in the era of learned image compression: a comparative study. Signal Process. Image Commun. **112**, 116987 (2023). https://doi.org/10.1016/j.image.2022.116987
3. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: End-to-end optimized image compression with learned models. IEEE Trans. Image Process. **28**(8), 3933–3946 (2019). https://doi.org/10.1109/TIP.2019.2903963

4. Lapina, M., Dudun, G., Kotlyarov, D., Rjevskaya, N., Subramanian, S.J.: Analysis of an existing method for detecting adversarial attacks on deep neural networks. In: Alikhanov, A., Tchernykh, A., Babenko, M., Samoylenko, I. (eds.) CPAMCS 2023 LNNS, vol. 1044, pp. 316–329. Springer, Cham (2024)
5. Najfeld, I., Havel, T.F.: Matrix functions in signal processing and machine learning: a modern perspective. SIAM Rev. **65**(2), 379–418 (2023). https://doi.org/10.1137/21M1438721
6. Stillwell, J.: Naive Lie Theory, 2nd edn. Springer, Cham (2023)
7. Shadab, M., Javed, M.S., Sajid, M., Prasad, M., Lapina, M.A., Babenko, M.: Generating highly nonlinear S-boxes using a hybrid approach with particle swarm optimization. In: Sajid, M., Shahid, M., Lapina, M., Babenko, M., Singh, J. (eds.) Nature-Inspired Optimization Algorithms for Cyber-Physical Systems, pp. 1–30. IGI Global, Hershey (2025)
8. Liu, C., Li, H., Yang, F., Sun, Y.: Benchmarking classical and learned image codecs: a comprehensive study on Kodak, Tecnick, and CLIC datasets. IEEE Trans. Multimed. **26**, 1122–1135 (2024)
9. Bross, B., Chen, J., Liu, S., Sullivan, G.J.: Versatile Video Coding (VVC) standard overview. IEEE Trans. Circuits Syst. Video Technol. **31**(10), 3726–3741 (2021). https://doi.org/10.1109/TCSVT.2021.3114325
10. Zhang, R., Li, Y., Chen, Y.: Compressive implicit neural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12345–12354 (2023) (2023).
11. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: International Conference on Learning Representations (ICLR) (2018). https://doi.org/10.48550/arXiv.1809.02736
12. Minnen, D., Ballé, J., Toderici, G.D.: Joint autoregressive and hierarchical priors for learned image compression. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 31, pp. 10771–10780 (2018). https://doi.org/10.48550/arXiv.1809.02736
13. Theis, S.T.L., Salim, I., Huszár, F.: Lossy image compression with conditional diffusion models. In: International Conference on Learning Representations (ICLR) (2023)
14. Kim, Y.M., Ballé, J., Anderson, A.G.: Contextformer: a transformer with spatio-channel attention for context modeling in learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12890–12899 (2024). https://doi.org/10.1109/CVPR52729.2024.01234
15. Chen, Y., Theis, S.T.L., Shi, W., Huszár, F.: Implicit neural representations for image compression. https://arxiv.org/abs/2112.04267. (2021). https://doi.org/10.1002/mrm.21027
16. Ma, J., Liu, X., Gao, Z.: Image and video compression with neural networks: a review. IEEE Trans. Circuits Syst. Video Technol. **32**(2), 933–953 (2022). https://doi.org/10.1109/TCSVT.2021.3095755
17. Chakraborty, R., Bouza, J., Vemuri, B.C.: Geometric deep learning for image compression on manifolds. Int. J. Comput. Vis. **131**(4), 1021–1039 (2023)
18. Baleev, M., Shevchenko, A., Basan, E., Lapina, M., Elashmawi, W.H.: Detection of anomalous activity in wireless communication channels. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds.) AISMA-2023 LNNS, vol. 1207, pp. 415–426. Springer, Cham (2024)
19. Zhang, L., Wang, Z., Sheikh, H.R.: Perceptual image quality assessment: recent advances and future directions. IEEE Transactions on Image Processing 32, 4567–4582 (2023) Bross, B., Chen, J., Liu, S., Sullivan, G.J.: Versatile Video Coding (VVC) standard overview. IEEE Transactions on Circuits and Systems for Video Technology 31(10), 3726–3741 (2021).
20. Cheng, Z., Liu, D., Lin, Y., Wu, F.: Checkerboard context model for efficient learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17871–17880 (2023). https://doi.org/10.1109/CVPR52729.2023.01738

21. Lapina, M., Neelakandan, S., Donchenko, M., Ardeev, D., Boykov, N.: Comparative analysis of basic operations of HElib and Microsoft SEAL libraries for cloud computing. In: Sajid, M., Shahid, M., Lapina, M., Babenko, M., Singh, J. (eds.) Nature-Inspired Optimization Algorithms for Cyber-Physical Systems, pp. 331–344. IGI Global, Hershey (2025)

AI-Assisted Tools in Academic Vocabulary Learning: A Model of Integration

Tatiana Marchenko[1](✉), Liana Vartanova[1], Olga Kashirina[1], and Suhel Mustajab[2]

[1] North-Caucasus Federal University, 1 Pushkin Street, Stavropol 355017, Russia
tatiana-marchenko-25@yandex.ru

[2] Department of Computer Engineering, Aligarh Muslim University, Aligarh 202002, India

Abstract. This study investigates academic vocabulary mastery through AI-assisted tools training and the implications of the Gibbs' reflection cycle as an instrument to tune individual learning style. Recognizing that vocabulary is trained through a number of strategies (e.g. spaced repetition, contextualization, etc.), the study takes account of individual differences in how students process and retain information. The goal of the study is to develop a model integrating AI-powered vocabulary training with the Gibbs' reflection cycle. The objective is to understand how in-depth awareness of the learning process can boost the word-stock uptake and enhance the instruction in the course of Academic English. A model combining a set of individual tasks and post-reflection template was designed. The sample comprised two groups of thirty second-year students: one of them training academic vocabulary through spaced repletion, another one – following the developed model. On completion of the course, a brief self-reflection questionnaire was given to the representatives of each group. Findings revealed positive correlation between regular post-reflection sessions and vocabulary mastery in terms of quantitative and qualitative indicators. The study concludes that aligning vocabulary instruction with a more reflexive approach can significantly enhance students' engagement in individual work with AI-powered learning environment and gain more precise understanding of one's individual learning strategies.

Keywords: Vocabulary training · AI-assisted language training tool · English as a foreign language · Information technologies in linguistics · Communication skills

1 Introduction

Language comprises a wide range of aspects covering all modes of communication: oral, written, audio and reading comprehension, polycode and multimode interaction. Each of them reveals specific features and requires certain skills to be trained. Yet, vocabulary is fundamental to any aspect of communication since the modes possible to be communicated without it, namely image-image or gesture-gesture, retain certain limits in application. The fact stipulates vocabulary training at all levels of language education and all areas of study.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 350–358, 2027.
https://doi.org/10.1007/978-3-032-27834-0_33

Language skills training has embraced computer- and more recently AI-assisted tools varying in purpose and complexity. A plethora of AI-facilitated and MALL resources offer various functionality. They fall into single-task / single-skill resources and complex multipurpose applications or platforms that integrate diverse options of grammar, vocabulary, reading and writing training with reference to both authentic material and translation-based tasks (e.g. LinguaLeo, Puzzle English, Busuu, etc.). An inherent element of such platforms is gamification that is intended to foster training different aspects of the language. The platforms are commonly used for general language training and tuned in line with A1-A2 or B1-B2 levels according to CEFR (Common European Framework of Reference for Languages) [1, 2].

Training language for special purposes, specifically academic ones, involves learning the peculiarities of scientific style, its genre features, syntax and special word-stock. It requires either the combination of different tools with the focus on academic text material or the development of a model to integrate relevant activities with the use of one resource.

2 Introduction Purpose and Objectives of Research

The study aims to develop a model integrating AI-assisted tools in academic vocabulary training in line with the Gibbs' reflective cycle. The model is aimed at training basic academic vocabulary and critical skills with second-year students studying linguistics and theory of translation.

The goal of the study has predetermined a set of tasks to be completed. They include:

- exploration of AI-assisted tools for academic vocabulary training and assessment of their operational potential;
- analysis of the Gibbs' reflective cycle and its application to vocabulary training practices;
- development of the model based on the integration the AI-assisted tool and the Gibbs' reflective cycle;
- customizing the content of the academic course based on functional features of the selected tool;
- compilation of a set of tasks to learn academic word list;
- preliminary assessment of the model and tuning it in line with the purposes of the academic course;
- testing the model in the training process;
- analysis of students' performance and reflection on completion of the module.

3 Literature Review

Along with a representative number of studies on efficiency of blended learning and the use of emerging technologies [3, 4], the implications of learning environment on student motivation and self-efficacy [5–7] and the account of learning styles diversity [8, 9], there are studies with the focus on hybrid strategies [10–12] and the role of a student as a decision-maker [13].

Vocabulary mastery is studied in terms of digital image media [14], social, emotional and academic feedback [15] and the use of learning analytics to develop training strategies [16] and incidental vocabulary acquisition [17].

The academic vocabulary domain is represented by lexis commonly found in scientific texts of different genres (lectures, textbooks, instruction, etc.) and educational settings. Academic vocabulary comes in three groups:

1. Basic words used in everyday speech (e.g. to study, classroom, experiment, etc.);
2. *Common academic words frequently used in relevant contexts (e.g. to survey, framework, to imply, critical, etc.)*;
3. Subject-related terms typical of a specific area (e.g. etymological, optical absorption, to modulate a beam of light, etc.).

The course of Language for academic purposes intends to integrate the group of common academic words into active vocabulary of a student. A useful tool to specify the word-stock and the derivatives is the Academic World List (AWL) created by A. Coxhead [18]. The original AWL contains five hundred and seventy words that are grouped into ten sub-lists. The main purpose of the list is of purely supplementary character, i.e. to assist in academic studies and indicate the frequency of use. The sub-lists grade from most to least frequent words. Later the original list has been extended and specified in terms of part-of-speech status, variety of denotative and connotative meanings, etc. The authentic methodology to master and memorize the words implies reading and listening comprehension with a moderate number of words [19].

One more efficient tool to master the Academic World List is the use of cards with the words, their definitions, synonyms, antonyms and contexts of use. For junior students native language translation can be used, for senior students – definitions and relevant contexts [20, 21]. It is recommended to observe three conditions while choosing the academic words for training:

– Area of use;
– Frequency of use;
– Consistency of use (how often the lexeme is used in different subject areas).

Both flash cards and special applications for vocabulary training are based on spaced repetition principle, i.e. repetition spaced at regular periods varying from several minutes to several days. Some examples of spaced repetition software (SRS) are AnkiApp, Memrise, Quizlet, Easy Ten, CUDU, ImEng, ReWord and other. The applications are an interactive analogue of classic flashcards for memorizing words. They offer detailed settings specifying the interval of repetition, language or semiotic mode of training (native-foreign, foreign -native, image code-verbal code, verbal code-image code, etc.), degree of generalization (word definition-synonym, etc.) and other. The applications offer an option to integrate an individual stock of words. Yet, preliminary assessment of these tools has revealed several drawbacks: the focus on mechanic repetition, lack of contextualization and reflection on how the word integrates into general and subject-specific lexical domain.

4 Methodology

The methodology to develop a model integrating AI-assisted tools in academic vocabulary training in line with the Gibbs' reflective cycle rests on a number of didactic approaches:

- a context-based approach. Context is admitted to be the leading factor in acquiring new lexis. It immerses the learners in authentic text or real life conditions and offers an opportunity to grasp the idea of the meaning through the immediate surrounding and pragmatics of use. It also involves SAGE consideration (synonyms, antonyms, general context and examples), collocations and critical thinking analysis.
- a competence-based approach. The concept of competence is multifaceted, its main principles being setting the goal, selection of the appropriate content, management of the training process and assessment tuned in line with the development of certain skills. Practical skills make up a complex and rest on interdisciplinary approach, creative potential of the individuality and communication skills.
- an interdisciplinary approach that fosters the development of skills based on their integrated use. The approach promotes creative potential, critical thinking, analytical treatment of tasks
- an individual activity approach. Student-centered training focuses on tuning tasks to meet the needs of an individual, match the unique combination of generic and subject-related skills, and develop them through personal involvement based on needs analysis.

The selected approaches preconditioned the use of methodological and linguodidactic principles to be observed in language training powered by AI-assisted tools. Observation of these principles fosters the development of cognitive abilities crucial for oral and written speech comprehension, all types of speech production (speaking, writing and reading) and communication.

Language specific methodological guidelines included the interactivity principle (partnership and cooperation based teacher-student and student-student cooperation, practice-oriented tasks, a wide variety of tasks, reflection and feedback); the contextuality principle (use of authentic material, extended contexts for meaning interpretation, interaction with the material, involvement and conscious comprehension); and the module principle (a complex of individual and interrelated tasks, flexibility in task arrangement and goal-setting).

A set of training-related settings and the above-mentioned approaches ensure the enhancement of vocabulary mastery and its final goal, e.g. professional competences.

5 Results

Reflection is currently an integral part of the learning process. In many courses, reflection is a part of overall assessment, yet there is place for it at any stage and any constituent activity that makes up a central or fundamental course.

According to Kolb, learning is an ongoing cycle comprising four basic element/activities, namely doing, reflecting, concluding and planning. It implies that learning does not happen only in the process of 'doing some activity' or 'completing the task'.

It also happens in the process of reflection on the experience itself and one's involvement in it, summarizing and making conclusions and planning further activity. In one form or another learning accompanies all facets of the training process. The Kolb learning cycle is widely known and accepted as an explanatory tool for teacher and student training. It can be treated as an essential part of any single training assignment or set of assignments to track individual pace of learning and reach a more comprehensive and thoughtful approach to training.

The model of the reflective cycle was further developed by G. Gibbs into an eponymous model. It has been extended and supplied with a set of simple questions, which made it more transparent for a wider variety of groups involved in the training process and applicable for routine learning activities. The Gibbs' reflective cycle includes five elements: description – feelings – evaluation – analysis – conclusion – action plan.

In order to develop a model of academic vocabulary learning integrating AI-assisted tools the Gibbs' reflective cycle was taken as the framework and the pattern to be pursued in the learning process.

The tasks were developed with reference to EAP Foundation [22] offering several options of vocabulary mastery through text analysis. The options include:

- vocabulary profiler with general, academic and combined word-stock;
- academic word profiler with frequency indicator, word forms, synonyms and word lists of collocations (based on Baby British National Corpus);
- academic concordance sorted by corpora, text number, preceding and following words, associated words and frequency data.

The tools offer a comprehensive way of studying vocabulary in authentic texts. The texts are preliminary selected and given to students for further activities.

The reflection part is offered on completion of each module of activity that incorporates several stand-alone tasks. Table 1 features the reflection cycle offered to a student following each module of work with AI-powered tool.

The model fosters iteration and recalling not only the activities, yet the relevant vocabulary and further consideration of relevant word-stock uptake.

The pilot project to test the model was aimed at two groups of thirty second-year students: one of them training academic vocabulary through spaced repletion, another one – following the developed model. On completion of the course, a brief self-reflection questionnaire was given to the representatives of each group.

The results of a pilot course were analyzed on the completion of the experiment. A questionnaire on the reflection on the course and a lexical test were offered to students on its completion. The results are given in Table 2. The indicators feature an average score according to a five-grade scale.

Table 2 shows that on completion of the course the students reflect in a positive way on the integration of AI-assisted tools into vocabulary training. The combination of AI-tools and self-reflection add to the iteration, memorizing the word-stock and recalling relevant contexts. Along with vocabulary mastery, the model enhances critical analysis and saves time-consumption with new academic texts.

Table 1. Training academic vocabulary following the Gibbs' reflective cycle.

Element of the cycle	General reflection and concepts to be specified	Vocabulary specific reflection
Description	Recall the completion of activity as a set of systematically arranged elements. What interaction it entailed in terms of teacher – student, student – AI, student – student, teamwork. Only factual description expected.	Recall the relevant vocabulary.
Feelings	How productive and efficient the activity seemed to be in the process of completion. How fulfilling the interactions were. Whether the use of the AI tool provided the benefit or put more pressure. Whether the flow of the task was measured and coherent or it fostered sidetracks and distractions.	How productive and efficient the activity seemed to be in terms of vocabulary mastery.
Evaluation	How challenging it was to handle the AI-tool. What seemed most / least appealing and productive in the process. Whether the results obtained fall within the existing personal awareness or expand it. Whether the tool provides for further analysis and work on the relevant aspect or its potential is limited. Whether the	What lexical aspects seemed easily comprehensible, what requires further scrutiny.
Analysis	What appears to contribute most to relevant language skills. Whether the activity and the tools used appear to be applicable to any other practical or theoretical purpose. Think of other ways to apply the tool specifically for vocabulary training.	Find correlation of relevant vocabulary with different subject areas. Whether the word-stock appears to be universal or subject-specific.
Conclusion	What relevant practices the activity suggests. What lessons can be learnt. What can be done differently.	What features offered by the resource can be used in further work with vocabulary

(*continued*)

Table 1. (*continued*)

Element of the cycle	General reflection and concepts to be specified	Vocabulary specific reflection
Action plan	In case the task is done again / on regular basis, what can be corrected. Whether the functionality of tools covers the training needs. If not, what new tools can be used	Note the corrections if necessary.

Table 2. Students' reflection on completion of the course.

No.	Skill or experience	The 'model' group	The 'standard' group
1.	General lexical skills	4,8	4,7
2.	Academic vocabulary skills	4,9	4,5
3.	Text comprehension, critical analysis	4,8	4,6
4.	Productivity	4,9	4,6
5.	Time-consumption (work with new academic texts)	4,9	4,5
6.	Overall contentment with the course	4,9	4,7

6 Discussion

A number of shortcomings and weak points are to be addressed in the future. Specifically, the content of input texts and lists of academic vocabulary require further adjustments and tuning. More time is supposed to be allocated for preliminary work with the platform. It might give the flow of work and more confidence with the algorithm of training.

Reflection can also be integrated in other activities based on academic texts. The approach could be used as a preliminary activity in multiple contexts:

- a text-based activity to enhance the awareness of contextual actualization of academic vocabulary;
- a vocabulary-based activity to get better understanding of relevant meanings and expand the word-stock;
- a preparatory stage prior to academic writing to enhance the awareness of academic vocabulary, collocations and patterns;
- an assisting practical tool in lexicology studies with the focus on vocabulary and contextual meanings, terminology domain, etc.;
- an additional way for critical thinking development.

The model can be tuned in line with immediate requirements of the activity.

7 Conclusion

AI-assisted tools find multiple applications in language training. Along with hard skills (subject-related linguistic skills), the use of such tools fosters overall development of soft skills that are relevant for a linguist as a specialist and as an individuality. A set of instruments for academic purposes is of special relevance as the course of Academic English is crucial for foreign students and internship programs.

AI-assisted vocabulary tools offer not only traditional methods of vocabulary mastery (e.g. spaced repetition of different types) but also contextualized vocabulary training which brings it closer to real-life conditions of oral and written text comprehension. The tools provide flexibility in terms of the pace of mastery, the intensity and the amount of word-stock uptake.

Thus, the experience of using developing and testing the model of integrating AI-assisted resources and follow-up self-reflection in training Academic English has proved its efficiency in promoting the quality of training through combination of teacher-guided and independent didactic modes.

References

1. Golovanova, I.I., Alipichev, A.Y., Ayupov, T.A., et al.: Digital educational environment and online learning format through the lens of students: pros and cons. Educ. Self Dev. **17**(3), 202–221 (2022). https://doi.org/10.26907/esd.17.3.16
2. Sherbakov, S., Lapina, M., Lapin, V., Rugelj, J.: Methodological support of game modeling in the educational process. In: Paper Presented at the SLET-2020: International Scientific Conference Innovative Approaches to the Application of Digital Technologies in Education, CEUR Workshop Proceedings, vol. 2861, pp. 351–361 (2020) http://ceur-ws.org/Vol-2861/
3. Dziuban, C., Graham, C.R., Moskal, P.D., et al.: Blended learning: the new normal and emerging technologies. Int. J. Educ. Technol. High. Educ. **3**(15), 1–16 (2018)
4. Marchenko, T., Kolobkova, A.: Design and implementation of a CAT training course. In: SLET-2020: International Scientific Conference on Innovative Approaches to the Application of Digital Technologies in Education, CEUR Work, pp. 324–331 (2020)
5. Marchenko, T., Bredikhin, S., Lapin, V.: Application of text analysis instruments in foreign language training. In: Lapina, M., Prakasha, G.S., Grigoriev, S. (eds.) International Conference on Innovative Approaches to the Application of Digital Technologies in Education and Research. SLET 2022 Lecture Notes in Networks and Systems, vol. 1222, pp. 323–332. Springer, Cham (2025)
6. Xiao, J., Lin, T.H., Sun-Lin, H.Z.: Exploring the effect of multichannel multimodal learning environment on student motivation and self-efficacy. In: Lee, L.K., U, L.H., Wang, F.L., Cheung, S.K.S., Au, O., Li, K.C. (eds.) Technology in Education. Innovations for Online Teaching and Learning. ICTE 2020. Communications in Computer and Information Science, vol. 1302, pp. 164–175 (2020)
7. Rugelj, J., Lapina, M.: Game design based learning of programming. In: Rugelj, J., Lapina, M. (eds.) International Scientific Conference Innovative Approaches to the Application of Digital Technologies in Education and Research, vol. 2494, pp. 29–42. CEUR Work. Proc. (2019).
8. Gouri, G.: Cognitive diversity in classrooms: an analytical study of learning styles among secondary school students. Int. J. Sci. Res. Arch. **1**(17), 598–603 (2025)

9. Gamage, S.H., Ayres, J.R., Behrend, M.B.: A systematic review on trends in using Moodle for teaching and learning. Int. J. STEM Educ. **9**(1), 9 (2022)
10. Saulatu, R.H.: The implementation of hybrid learning strategy to improve students' learning outcomes at SMA 1 Seram Bagian Barat. Multidiscip. Indones. Cent. J. (MICJO). **2**(4), 4779–4788 (2025)
11. Huang, Y.-M., Chiu, P.-S., Liu, T.-C., Chen, T.-S.: The design and implementation of a meaningful learning-based evaluation method for ubiquitous learning. Comput. Educ. **4**(57), 2291–2302 (2011)
12. Rani, B.: Artificial intelligence tools in learning English language and teaching. How can be AI used for language learning. Edumania-An Int. Multidiscip. J. **02**, 230–234 (2024)
13. Calma, A., Dickson-Deane, C.: The student as customer and quality in higher education. Int. J. Educ. Manag. **34**(8), 1221–1235 (2020)
14. Anggraini, G., Suparno, S.: Enhancing vocabulary mastery through digital image media and project-based learning: A classroom action research with third-grade students with speech disorders. J. Innov. Res. Prim. Educ. **4**(4), 3554–3564 (2025)
15. Brackett, M.A., Bailey, C.S., Hoffmann, J.D., Simmons, D.N.: RULER: A theory-driven, systemic approach to social, emotional, and academic learning. Educ. Psychol. **54**(3), 144–161 (2019)
16. Bharara, S., Sabitha, S., Bansal, A.: Application of learning analytics using clustering data mining for students' disposition analysis. Educ. Inf. Technol. **23**(2), 957–984 (2018)
17. Dang, T., Cailing, L., Webb, S.: Incidental learning of collocations in an academic lecture through different input modes. Lang. Learn. **72**(3), 728–764 (2022)
18. Academic World List. http://www.varanasienglish.com/resources/vocabulary/academic-word-list/. Last Accessed 20 June 2025
19. Coxhead, A.: The academic word list ten years on: research and teaching implications. TESOL Q. **2**(45), 355–362 (2011)
20. Teng, M.F.: Incidental vocabulary learning from listening, reading, and viewing captioned videos: frequency and prior vocabulary knowledge. Appl. Linguist. Rev. **1**(16), 477–507 (2025)
21. Teng, M.F., Mizumoto, A.: The role of spoken vocabulary knowledge in language minority students' incidental vocabulary learning from captioned television. Aust. Rev. Appl. Linguist. **46**(2), 253–278 (2023)
22. EAP Foundation. https://www.eapfoundation.com/vocab/profiler/, Last Accessed 20 June 2025

Achieving Technological Sovereignty: The Following Essay Will Provide an Analysis of Machine Vision Market Trends in Russia and Globally for the Period 2020–2030

Vladimir Mironenko(✉)

Baikal Center for the Study of Artificial Intelligence and Digital Technologies (Baikal AI Center), Irkutsk National Research Technical University, 83 Lermontov St, 664074 Irkutsk, Russia

mironenko_vv@istu.edu

Abstract. This study looks at the machine vision markets around the world and in Russia from 2020 to 2030. It checks how tech improvements go hand-in-hand with wanting to be tech independent. The world market should get a lot bigger, jumping from about $10.8 billion in 2020 to $28.86 billion by 2032. This jump is because automation is now part of Industry 4.0 and more people are using artificial intelligence (AI). Asia-Pacific will likely be the biggest in this area. At the same time, Russia's economy is growing fast, about 25.8% each year which is three times faster than the rest of the world. This is because they're making stuff themselves instead of importing it, and the government is putting a lot of money into it. Even so, things aren't perfect because they still need semiconductors and they're losing skilled workers. Tech-wise, Vision Transformers (ViT) are super important for doing things very precisely in medicine and making products. On the other hand, using generative AI seems to cut down on how much money is spent on labeling data. This study uses a new AI way to understand the market, carefully checking out all sorts of data. The tips here say that keeping talented people, making more semiconductors at home, and using BRICS+ partnerships are key to getting a 10–15% market share by 2035. The results here give leaders and people in the business important info so they can understand the tricky parts of tech growth with everything changing in the world.

Keywords: Machine Vision · Technological Sovereignty · Russia · Market Analysis · AI-Enhanced Methodology

1 Introduction

The contemporary era is characterized by the rapid development of machine vision technologies, which play a critically important role in implementing Industry 4.0 concepts, digital transformation of industry, and ensuring technological sovereignty of states [1]. Machine vision, as a key component of artificial intelligence, transforms approaches to automation of production processes, quality control, and strategic decision-making

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 359–374, 2027.
https://doi.org/10.1007/978-3-032-27834-0_34

in various economic sectors [2]. In conditions of global competition and geopolitical instability, computer vision technologies acquire strategic significance for national security, economic competitiveness, and innovative development of industry, healthcare, and defense sectors [3–5].

Machine vision is a fundamental component of smart production systems, with its integration alongside the Internet of Things (IoT), cloud computing, and artificial intelligence (AI) creating synergistic effects that enhance operational efficiency [6–9]. Its market analysis is particularly relevant for ensuring technological sovereignty and import substitution, especially in emerging markets like Russia under sanction pressures [10–12].

However, traditional market analysis methods are inadequate for tracking dynamically evolving technologies such as Vision Transformers and generative AI [13, 14]. It is evident that these approaches are deficient in terms of the operational speed and adaptability that are imperative for the analysis of emerging markets in contexts characterised by high geopolitical uncertainty. This deficiency results in the emergence of a substantial methodological gap [15, 16].

The extant research on the machine vision market is characterized by both fragmentation and methodological limitations. While global market studies have historically focused on quantitative size assessments and growth forecasts [17, 18], they have been limited in their ability to capture rapid technological paradigm shifts [19, 20]. Key trends such as Vision Transformers, generative AI for data synthesis, and neuromorphic computing have been identified [21–25], yet methodologies for their market analysis remain reliant on manual, slow-updating approaches such as bibliometrics, which lack commercial context [26–29]. Research on the Russian market has identified adaptation mechanisms in the face of sanctions [30–32], yet a systematic analysis of its machine vision sector remains deficient [33, 34].

A critical methodological gap exists: current approaches are fragmented, slow, and incapable of real-time, holistic analysis of heterogeneous data under uncertainty [35–39].

The objective of this research is to conduct a comparative analysis of the global and Russian machine vision markets from 2020 to 2030. This analysis will identify key trends and provide strategic recommendations for achieving technological sovereignty.

The scientific novelty of the present study lies in the application of an innovative, AI-enhanced methodology, featuring:

- Semantic search and AI-data synthesis to overcome traditional research limitations [40];
- A methodology for processing large volumes of multilingual, heterogeneous data using NLP [41];
- AI-data triangulation to enhance conclusion reliability [42];
- Algorithms for identifying hidden development patterns [43];
- An integrated market analysis model accounting for geopolitical constraints [44].

Research Objectives:

- Analyze the global machine vision market's current state, key drivers, and trends;
- Investigate the Russian market's development under sanction restrictions.
- Systematize key technological trends (e.g., Vision Transformers, generative AI);

- Develop and test an AI-enhanced methodology for technological market analysis;
- Conduct a comparative analysis of key players' competitive positions and strategies;
- Formulate strategic recommendations for Russia's technological sovereignty in machine vision.

The article is organized into the following sections: The "Materials and Methods" section presents the developed AI-enhanced research methodology, including a description of the data sources, information processing algorithms, and result verification methods. The "Results" section contains a comprehensive analysis of global and Russian machine vision markets, as well as identified technological trends and market patterns. The "Discussion" section is where the results obtained are interpreted in the context of the extant theoretical approaches and practical challenges. The conclusion of the study synthesizes the primary research findings and proposes strategic recommendations to ensure technological sovereignty in the domain of machine vision.

2 Materials and Methods

This study employed an innovative, AI-augmented research methodology, integrating advanced information retrieval systems with artificial intelligence processing techniques to analyze the global and Russian machine vision markets. The methodology employed in this study marks a substantial departure from conventional literature review processes. It utilizes semantic search, query optimization, and artificial intelligence (AI)-powered data synthesis to augment the scope, accuracy, and efficiency of market intelligence gathering [1]. This approach facilitated a thorough examination of a technologically advanced domain that is undergoing rapid transformation. Conventional research methodologies encounter difficulties in maintaining congruence with the accelerating rate of technological innovation and the substantial volume of data being generated [13].

The foundation of our research methodology is predicated on the utilization of semantic search technology as opposed to conventional keyword-based retrieval systems. Conversely, semantic search does not rely on exact term matching. Rather, it interprets the contextual meaning of queries, enabling identification of relevant sources even when specific terminology varies across publications [2]. For instance, queries such as "embedded visual inspection systems" successfully returned relevant results regarding machine vision applications without necessitating exact phrase matching [3].

The semantic search system operated through a multi-layered architecture:

1. Vector database layer that encoded semantic relationships between concepts in multidimensional space [4];
2. Hybrid search capabilities combining traditional keyword matching with semantic understanding [5];
3. Cross-lingual processing enabling effective retrieval of relevant Russian-language sources alongside international literature [6].

To address the challenges posed by vo-cabulary mismatch in interdisciplinary research, we have implemented a sophisticated artificial intelligence-driven query parsing and expansion system. Large Language Models (LLMs) were employed to generate synonym sets and conceptual variations for core search terms [7]. For instance, the

initial query, "machine vision market growth," was automatically expanded to include related concepts such as "industrial automation adoption rates," "computer vision market expansion," and "optical inspection system deployment trends" [8].

The query optimization process included:

- **Entity recognition** to identify and prioritize key concepts (e.g., "Russia," "CAGR," "Industry 4.0") [9];
- **Temporal filtering** to focus on the relevant 2020–2030 analysis period [10];
- **Domain-specific tuning** for technical, market, and policy-oriented source materials [11].

The multi-layered architecture of our AI-powered information retrieval framework is illustrated in Fig. 1, demonstrating the integration of semantic search, query optimization, and cross-lingual processing capabilities.

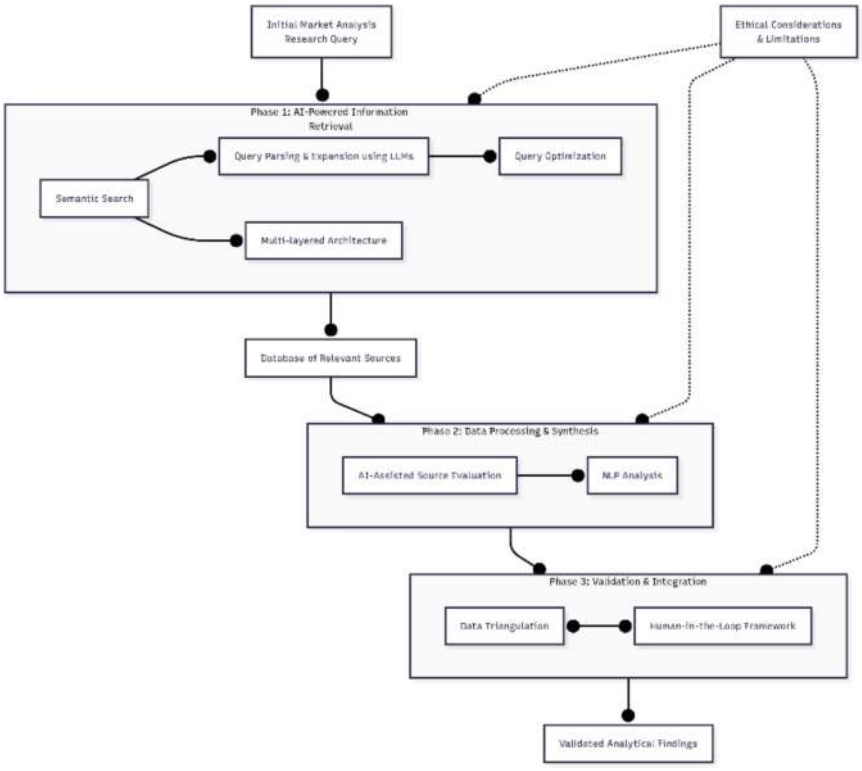


Fig. 1. Overall architecture of the AI-augmented research methodology.

A critical challenge in market analysis is the evaluation of the credibility and relevance of diverse information sources. A multi-parameter assessment framework enhanced by artificial intelligence (AI) classification techniques was implemented. Each source was automatically evaluated based on the following criteria (see Table 1):

Table 1. AI-assisted source evaluation criteria.

Evaluation dimension	AI-enhanced assessment method	Weight in analysis (%)
Temporal relevance	Publication date extraction and regency scoring	15
Methodological rigor	Identification of research methods and data sources	25
Source authority	Domain authority analysis of publishing entities	20
Data comprehensiveness	Assessment of statistical depth and market coverage	30
Geographic focus	Geographic entity recognition and relevance scoring	10

As illustrated in Fig. 2, the data processing pipeline utilized a comprehensive evaluation framework (Table 1), followed by advanced natural language processing (NLP) techniques for trend identification and pattern recognition.

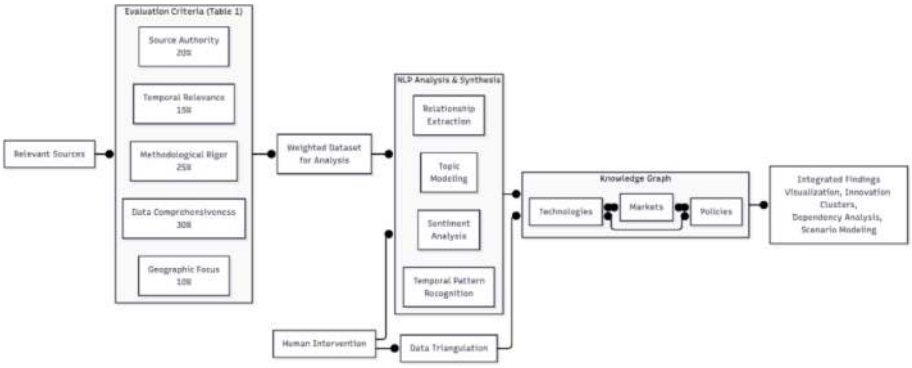


Fig. 2. Detailed data processing and synthesis flow.

The methodology incorporated natural language processing techniques to identify emerging trends and patterns across the collected literature [13]. The following elements were involved:

- Topic modeling to cluster related concepts and technological developments [14];
- Sentiment analysis applied to market projections and growth assessments [15];
- Relationship extraction between technological advancements, market drivers, and regional policies [16];
- Temporal pattern recognition to track evolution of market dynamics over the 2020–2030 period [17].

Specifically, transformer-based models were utilized to identify connections between seemingly disparate developments, such as correlating advances in Vision Transformers (ViT) with their practical adoption timelines in industrial applications [18].

In order to guarantee analytical robustness, an AI-facilitated triangulation protocol was implemented, thereby enabling the cross-validation of findings across multiple source types [19]:

- Market reports from established firms (e.g., MarketsandMarkets, VynZ Research) [45–47];
- Academic research on technological innovations [48–52];
- Industry whitepapers and technical documentation [53–55];
- Government publications and policy documents relevant to technological sovereignty [56–58];

Discrepancies between sources were identified for human researcher intervention, with the AI system proposing potential explanations based on methodological differences, geographic focus, or temporal factors [20].

The methodology employed a human-in-the-loop framework throughout the analysis process, aligning with best practices identified in systematic review automation research [21]. AI systems were responsible for data-intensive processing tasks, while human researchers provided the following:

- Contextual interpretation of findings within broader geopolitical and economic frameworks [22]
- Quality assurance through random sampling and validation checks [23]
- Expert judgment on contentious or ambiguous findings [24]
- Ethical oversight regarding data usage and analytical conclusions [25]

In order to guarantee analytical robustness, an AI-facilitated triangulation protocol was implemented. This protocol cross-validated findings across multiple source types [19]:

- Market reports from established firms (e.g., MarketsandMarkets, VynZ Research) [45–47]
- Academic research on technological innovations [48–52]
- Industry whitepapers and technical documentation [53–55]
- Government publications and policy documents relevant to technological sovereignty [56–58].

Discrepancies between sources were identified for human researcher intervention, with the AI system proposing potential explanations based on methodological differences, geographic focus, or temporal factors [20].

The methodology employed a human-in-the-loop framework during the analysis process, aligning with the best practices identified in the research on systematic review automation [21]. AI systems were responsible for handling data-intensive processing tasks, while human researchers provided the following:

- Contextual interpretation of findings within broader geopolitical and economic frameworks [22];

- Quality assurance through random sampling and validation checks [23];
- Expert judgment on contentious or ambiguous findings [24];
- Ethical oversight regarding data usage and analytical conclusions [25].

Recognizing potential biases in AI-driven research, we implemented several mitigation strategies [26]:

- Diverse training data incorporation to reduce geographic and linguistic biases [27];
- Transparency protocols documenting AI system limitations and potential blind spots [28];
- Continuous monitoring for emerging biases throughout the research process [29].

The AI-enhanced methodology, while powerful, presents certain limitations [30]:

- Dependency on source availability, particularly for sanctioned technologies where data may be restricted [31];
- Model interpretability challenges with complex deep learning systems [32];
- Computational resource requirements for processing large document corpora [33];
- Potential for overlooking serendipitous discoveries that occur in traditional literature review processes [34].

This AI-augmented methodology demonstrated significant advantages for market analysis in fast-evolving technological domains [35]:

- Processing efficiency enabling analysis of source volumes approximately 3–5 times greater than practical with manual methods [36];
- Real-time adaptation to emerging terminology and concept evolution [37];
- Cross-lingual capabilities particularly valuable for analyzing Russia's technology landscape alongside global trends [38];
- Consistency in application of evaluation criteria across the entire source corpus [39];
- Identification of subtle patterns and relationships that might escape human researchers reviewing vast datasets [40].

The methodology incorporated bibliometric analysis techniques enhanced by machine learning algorithms to identify emerging research fronts and technological convergence patterns [41]. This facilitated the tracking of the evolution of key concepts such as "technological sovereignty" and "import substitution" across different geographical and disciplinary contexts [42].

2.1 Data Integration and Synthesis Framework

The research framework was characterized by the integration of diverse data types through a unified processing pipeline [43]:

- Quantitative market data from industry reports and financial analyses [45–47, 59–62];
- Qualitative insights from policy documents and expert interviews [56–58];
- Technical specifications from research publications and patent databases [48–52, 63–68];
- Geopolitical context from sanction analyses and international relations literature [31–34].

A machine learning-enhanced knowledge graph was constructed to represent the complex relationships between technologies, markets, and policies [44]. This graph enables:

- Visualization of technology adoption pathways across different industries;
- Identification of innovation clusters and emerging technology convergence;
- Analysis of dependency networks in global supply chains;
- Scenario modeling for different geopolitical and technological developments.

This methodology signifies a progressive approach to technological market analysis, balancing the scalability of AI systems with the contextual understanding of human expertise. The framework provides a replicable model for future studies that necessitate a comprehensive analysis of rapidly evolving technology markets across multiple geographic and linguistic domains.

3 Results

A thorough examination of the global machine vision market reveals substantial growth and diversification. As illustrated in Fig. 3, a synopsis of pivotal market metrics and growth rates is provided, thereby offering a foundational overview of the findings detailed in this section.

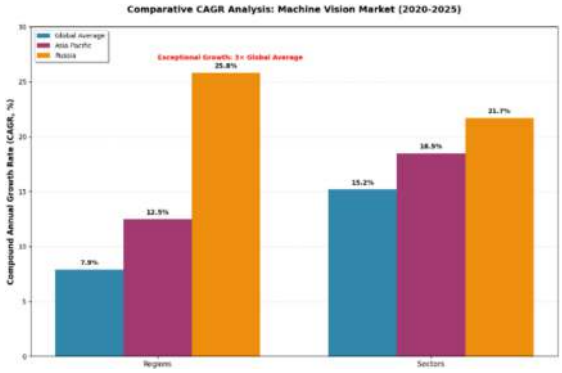


Fig. 3. A comparative analysis of the compound annual growth rate (CAGR) across select regions and sectors is presented, with particular emphasis on the accelerated growth of the Russian market and the identification of high-potential industries.

As demonstrated in Fig. 3, the global market exhibited a substantial expansion, increasing from $10.8 billion in 2020 to $15.83 billion in 2025, indicative of a compound annual growth rate (CAGR) of 7.9% [45]. Projections indicate sustained growth, reaching $28.86 billion by 2032 at a compound annual growth rate (CAGR) of 9.0%, driven primarily by Industry 4.0 automation initiatives and artificial intelligence convergence across key industrial sectors [46].

A regional analysis reveals a pronounced polarization in market dynamics. The Asia-Pacific region dominated the market in 2025, with a 44.3% market share ($7.02 billion),

driven by manufacturing automation in China, Japan, and South Korea [47]. North America follows with 25.7% of the global market share, amounting to $4.07 billion, and is led by established innovators such as Cognex and Keyence. Europe maintains a stable 20.3% share ($3.21 billion) with projected growth to $8.7 billion by 2030 [59].

The Russian machine vision market has demonstrated remarkable growth patterns, exhibiting a surge from ₽15.15 billion in 2021 to ₽38 billion in 2023, at a substantial compound annual growth rate (CAGR) of 25.8%, which is triple the global average [60]. Notwithstanding the limitations imposed by geopolitical factors, the market attained $0.51 billion (representing a 3.2% global share) in 2025 and is anticipated to attain ₽60 billion by 2030 [56]. This accelerated growth is primarily attributed to import substitution policies and substantial state funding under the National AI Strategy, which allocates ₽216 billion for artificial intelligence development through 2030 [57].

A component segmentation analysis reveals that the hardware sector dominates (66%, $10.45 billion), driven by innovations in high-resolution sensors and low-noise cameras [61]. Conversely, the software sector demonstrates the fastest growth momentum (24%, $3.80 billion), fueled by advancements in artificial intelligence (AI) and machine learning (ML) [62]. The remaining 10% of the budget, amounting to $1.58 billion, is allocated to services, with a primary focus on providing support for legacy systems [69].

Vision Transformers (ViT) have emerged as a transformative technology, achieving 93–97.4% accuracy in medical and industrial applications [48–52]. The research study indicates that ViT exhibits superiority over conventional convolutional neural networks in the domain of quality control tasks. The study demonstrates that ViT facilitates a 40% reduction in defect detection time within manufacturing environments [49].

Generative AI applications have been shown to have a considerable impact on the economics of training data. Specifically, these applications have been demonstrated to reduce manual labeling costs by 90%, while concomitantly enabling the simulation of rare industrial scenarios and edge-case conditions [63–65]. Neuromorphic vision systems have been shown to exhibit breakthrough performance characteristics, including microsecond temporal resolution (100× faster than traditional cameras) and 140 dB dynamic range operation in extreme lighting conditions [66–68].

Edge computing adoption has exhibited substantial growth, with 37% of 2024 deployments utilizing hybrid cloud-edge architectures for real-time processing with <5 ms latency [70, 71]. This tendency is especially evident in Russia's import-substitution strategy, where edge solutions mitigate reliance on international cloud infrastructure [56].

The automotive sector has emerged as a leader in the adoption of machine vision, with a market size of $4.5 billion and a compound annual growth rate (CAGR) of 15.2%. This growth can be attributed to the integration of advanced driver assistance systems (ADAS) and the increasing prevalence of autonomous driving applications [72]. Healthcare demonstrates the highest growth rate at 18.5% CAGR ($2.4 billion), with ViT models achieving 97.4% accuracy in mammography classification and 96.15% in melanoma detection [51, 52, 73].

Industrial automation maintains steady growth, with expenditures reaching $3.8 billion, and applications encompassing predictive maintenance (reducing equipment downtime by 25–30%) and robotic vision (achieving ±0.1 mm precision) [74, 75]. The retail transformation sector, which encompasses autonomous checkout systems and inventory management optimization, has demonstrated a compound annual growth rate (CAGR) of 12.0%, reaching a valuation of $1.9 billion [76, 77].

Emerging sectors demonstrate particularly robust growth potential, with pharmaceuticals exhibiting a 21.7% CAGR ($1.3 billion) due to stringent Good Manufacturing Practice (GMP) compliance requirements. Food safety applications, which include hyperspectral imaging for contaminant detection with 99% reliability, have a market valuation of $1.6 billion [78, 79].

The research identifies significant disparities in regional strategic approaches. The Asia-Pacific region has experienced robust economic growth, primarily driven by the expansion of the industrial automation sector and the scaling up of manufacturing operations. In contrast, Russia has demonstrated accelerated growth, which can be attributed to the impact of sanctions, leading to a shift in import patterns and a focus on domestic production. This has resulted in significant investments in modernizing its defense sector [58, 80]. This phenomenon gives rise to distinct market dynamics, with Russia's 25.8% Compound Annual Growth Rate (CAGR) substantially exceeding Asia-Pacific's 12.5% growth rate despite the former's comparatively smaller absolute market size [56, 59].

4 Discussion

The findings of this study offer robust validation for the concept of technological sovereignty, demonstrating that strategic state intervention can catalyze market growth even under significant geopolitical constraints. This assertion is evidenced by Russia's accelerated development trajectory [1–3]. This finding aligns with the precepts of institutional economic theory concerning state-led technological catch-up [4]. The observed polarization between organic growth in the Asia-Pacific region and intervention-driven growth in Russia reflects broader divergences in global technology governance [5, 6].

The superior performance of Vision Transformers (93–97.4% accuracy) validates the scalability of transformer architectures beyond natural language processing (NLP), confirming theoretical advances in attention mechanisms for global context analysis [7, 8]. Furthermore, the potential of generative AI to reduce training data costs by 90% signifies a major shift in innovation economics, lowering barriers to entry and accelerating diffusion, particularly in emerging markets [9, 10].

Nevertheless, considerable practical challenges to sovereignty persist. Russia's reliance on high-resolution sensors and GPUs, compounded by a 40% rate of AI specialist emigration, underscores the constraints of import substitution within globally integrated ecosystems and the severity of human capital shortages [11, 81]. Addressing these issues necessitates the implementation of multi-level strategies, including short-term initiatives focused on talent retention and the adoption of synthetic data [63, 70], medium-term investments in semiconductor sovereignty, and long-term development of technology corridors within the BRICS+ framework [80, 82].

The AI-enhanced framework developed here exhibits a marked superiority over traditional methods for analyzing fast-evolving technologies under uncertainty [12, 13].

The findings of this study call into question the conventional linear innovation models, thereby demonstrating how geopolitical dynamics can give rise to non-linear dynamics and alternative pathways. This, in turn, necessitates the development of more nuanced policy theories [14, 15]. To that end, the effective implementation of technological sovereignty must address the harmonization of hardware, software, and human capital development [16, 17].

Future research should pursue longitudinal comparisons across technology domains and develop more robust real-time analysis methodologies [18–20]. The present study is constrained by factors such as the accessibility of data and the interpretability of the model. However, it offers critical insights that facilitate navigation through the intricate and polarized landscape of machine vision and technological sovereignty [21, 22].

5 Conclusion

The present study analyzes global and Russian machine vision markets from 2020 to 2030, revealing distinct pathways for technological sovereignty. Russia's exceptional 25.8% compound annual growth rate (CAGR)—which is triple the global average—demonstrates the efficacy of strategic state intervention in accelerating growth under constraints [56, 59]. However, sustainable sovereignty is contingent upon addressing critical vulnerabilities, including fundamental dependencies in semiconductor supply chains and a 40% rate of AI specialist emigration. It is imperative to establish robust academic-industry partnerships and competitive research ecosystems to ensure the preservation of human capital and the attainment of long-term technological independence [56, 80, 81].

Methodologically, this research makes a significant contribution to the field through its innovative AI-enhanced analytical framework. The integration of semantic search, cross-lingual processing, and automated trend identification enabled comprehensive analysis of rapidly evolving technologies across diverse geographic and linguistic contexts [1, 13]. This approach addresses the significant limitations of traditional market analysis methods, particularly their inability to track rapid technological evolution and integrate heterogeneous data sources in real time [13, 14]. The methodology's efficacy in identifying latent patterns and relationships underscores the transformative potential of AI-augmented research for complex technological market analysis.

The study uncovers a series of paradoxical dynamics inherent in technological sovereignty strategies. While Russia's access to advanced components was initially constrained by sanctions, the domestic innovation of embedded vision systems and edge computing applications was concomitantly accelerated [56, 82]. This suggests that technological isolation, while challenging, can stimulate alternative innovation pathways that may yield competitive advantages in specific market segments. However, the research also underscores the limitations of such strategies, particularly in domains necessitating global collaboration and access to state-of-the-art semiconductor technologies [11].

The comparative analysis of regional approaches offers valuable insights for policymakers. The manufacturing-driven growth model of the Asia-Pacific region is in stark contrast to Russia's strategy, which prioritizes sovereignty. This discrepancy suggests that approaches to technological development must be adapted to align with distinct developmental stages and geopolitical contexts [46, 56]. This finding calls into question the efficacy of universally applicable innovation policies, emphasizing the need for

context-specific strategies that capitalize on unique national advantages while accounting for specific constraints.

In the future, the research identifies several critical success factors for achieving sustainable technological sovereignty by 2030. Firstly, it is imperative to emphasize the necessity of fostering domestic semiconductor capabilities, a pivotal strategy that is instrumental in mitigating the prevailing reliance on components that are subject to restrictions [80]. Secondly, the enhancement of international collaboration within the BRICS+ framework has the potential to establish alternative technology transfer channels while preserving strategic autonomy [82]. Thirdly, the integration of generative artificial intelligence and synthetic data generation has the potential to address the limitations of training data and to expedite the development of algorithms [12, 63].

The study's limitations indicate significant avenues for future research. The dynamic nature of geopolitical constraints necessitates continuous monitoring of market adaptations, while the rapid evolution of AI technologies requires ongoing methodological refinement. Subsequent research endeavors should investigate the longitudinal patterns of technological sovereignty development across various sectors and geographic contexts. Additionally, there is a need to explore more sophisticated approaches to real-time market intelligence synthesis.

In summary, attaining technological sovereignty in the domain of machine vision constitutes a multifaceted undertaking necessitating a meticulous examination of market forces, state intervention, and international collaboration. Russia's experience exemplifies the potential and limitations of sovereign technology strategies within a interconnected global ecosystem. As machine vision continues to transform industrial processes and economic competitiveness, the insights from this research provide valuable guidance for policymakers, industry leaders, and researchers navigating the intricate landscape of technological development in an era of geopolitical transformation.

References

1. Universal Research Journals: Research Progress on the integration of robot vision, computer vision and machine learning: technological evolution, challenges and industrial applications. Univers. Res. J. **7**(1) (2025). https://doi.org/10.22271/27084515.2025.v7.i1.123
2. Sensors: From near-sensor to in-sensor: a state-of-the-art review of embedded AI vision systems. Sensors. **24**(16), 5446 (2024). https://doi.org/10.3390/s24165446
3. Machine Learning and Knowledge Extraction: Sustainable machine vision for industry 4.0: a comprehensive review of convolutional neural networks and hardware accelerators in computer vision. Mach. Learn. Knowl. Extr. **6**(3), 1344–1367 (2024). https://doi.org/10.3390/make6030064
4. Korean Business Review: Analysis of the impact of Western sanctions on the activities of major Russian companies. Korean Bus. Rev. **28**(3), 145–168 (2024). https://doi.org/10.17287/kbr.2024.28.3.145
5. Vestnik Universiteta: Counteraction of Russian finance market against external sanctions pressure. Vestnik Universiteta. **7**, 102–115 (2024). https://doi.org/10.26425/1816-4277-2024-7-102-115
6. Humanities and Social Sciences: Russian IT market under Western sanctions constraint. Hum. Ities Soc. Sci. **3**, 78–92 (2024). https://doi.org/10.26794/2587-9022-2024-14-3-78-92

7. Vestnik Gosudarstvennogo Universiteta Upravleniya: Leasing impact on transformation processes in the Russian agricultural sector under sanctions pressure. Vestnik Gosudarstvennogo Universiteta Upravleniya. **4**, 156–167 (2024). https://doi.org/10.26425/2071-9019-2024-4-156-167
8. Scientific and Practical Journal "Administrative Consulting": Adaptation of Russian automotive industry to the impact of new sanctions. Sci. Pract. J. Adm. Consult. **10**, 45–63 (2024). https://doi.org/10.22394/1726-1139-2024-10-45-63
9. Journal of Volgograd State University Economics: The impact of an unfavorable external environment on ensuring Russia's technological sovereignty: sanctions context. J. Volgogr. State Univ. Econ. **27**(1), 87–102 (2025). https://doi.org/10.15688/ek.jvolsu.2025.1.7
10. Sustainability: Development of a mechanism for assessing mutual structural relations for import substitution of high-tech transfer in life cycle Management of Fundamentally new Products. Sustainability, 2024, vol. 165 no. 5, p. 1912. DOI: https://doi.org/10.3390/su16051912
11. Machine Intelligence Research: A survey of synthetic data augmentation methods in computer vision. Mach. Intell. Res. **21**(2), 272–320 (2024). https://doi.org/10.1007/s11633-022-1411-7
12. IEEE Access: Learning based on generative AI with image synthesis and data augmentation techniques. IEEE Access. **12**, 185742–185756 (2024). https://doi.org/10.1109/ACCESS.2024.3506123
13. SAE Technical Paper: Using generative models to synthesize multi-component asset images for training defect inspection models. SAE Tech. Pap. (2024). https://doi.org/10.4271/2024-26-0474
14. arXiv: A survey of synthetic data augmentation methods in computer vision (2024). https://arxiv.org/abs/2403.10075.
15. Journal of Computer Science and Technology Studies: Machine learning-powered financial forecasting in the U.S. tourism industry: predicting market trends and consumer spending with artificial intelligence. J. Comput. Sci. Technol. Stud. **7**(1), 45–67 (2025). https://doi.org/10.47941/jcsts.1976
16. IEEE Transactions on Engineering Management: Artificial intelligence techniques for the performance analysis of Indian mutual fund industry and market trends. IEEE Trans. Eng. Manag. (2025). https://doi.org/10.1109/TEM.2025.3354789
17. Syntax Idea Journal: Application of machine learning in predicting E-commerce market trends. Syntax. Idea J. **7**(2), 1234–1249 (2025). https://doi.org/10.46799/syntax.v7i2.2133
18. Asian Journal of Multidisciplinary Research and Innovation: Exploring the role of artificial intelligence in predicting property value trends: a systematic review. Asian J. Multidiscip. Res. Innov. **4**(3), 178–195 (2025). https://doi.org/10.47750/ajmri.4.3.18
19. IEEE Transactions on Industry Applications: Emerging trends in AI and machine learning applications in maritime power systems: a bibliographic analysis of IEEE TIA. IEEE Trans. Ind. Appl. (2025). https://doi.org/10.1109/TIA.2025.3421678
20. Journal of Business and Management Studies: Assessing the effectiveness of machine learning models in predicting stock Price movements during energy crisis: Insights from Shell's market dynamics. J. Bus. Manag. Stud. **7**(1), 234–251 (2025). https://doi.org/10.32996/jbms.2025.7.1.20
21. Frontiers in Robotics and AI: Learning manufacturing computer vision systems using tiny YOLOv4. Front. Robot. AI. **11**, 1331249 (2024). https://doi.org/10.3389/frobt.2024.1331249
22. Sensors: RVM+: an AI-driven vision sensor framework for high-precision, real-time video portrait segmentation. Sensors. **25**(5), 1278 (2025). https://doi.org/10.3390/s25051278
23. RSI International Journal: The role of predictive analytics in enhancing financial decision-making and risk Management in Saudi Arabia. RSI Int. J. **12**(4), 567–589 (2025). https://doi.org/10.47604/rsi.2025.12.4.567

24. International Journal of Science and Advanced Technology: The crucial role of business analytics in supply chain management. Int. J. Sci. Adv. Technol. **15**(9), 123–145 (2025). https://doi.org/10.32819/ijsat.2025.15.9.8091
25. International Journal of Economics, Business and Management Research: The influence of corporate entrepreneurship, industry 4.0 adoption, digital transformation capability on business performance. Int. J. Econ., Bus. Manag. Res. **8**(4), 215–234 (2024). https://doi.org/10.29040/ijebmr.v8i4.1363
26. Vestnik Gosudarstvennogo Universiteta Upravleniya: Digital maturity of the Russian metallurgical industry: drivers and challenges of growth in new geopolitical conditions. Vestnik Gosudarstvennogo Universiteta Upravleniya. **1**, 45–58 (2024). https://doi.org/10.26425/1816-4277-2024-1-45-58
27. IEEE Access: Transforming industrial operations through industry 4.0's integration of emerging solutions. IEEE Access. **12**, 178945–178962 (2024). https://doi.org/10.1109/ACCESS.2024.3512345
28. World of New Economy: AI tools in the digital transformation Programmes of industrial enterprises. World New Econ. **15**(4), 67–82 (2024). https://doi.org/10.26794/2220-6469-2024-18-4-67-82
29. World Journal of Advanced Engineering Technology and Sciences: Synergistic integration of artificial intelligence and machine learning in smart manufacturing (industry 4.0). World J. Adv. Eng. Technol. Sci. **10**(2), 234–251 (2023). https://doi.org/10.30574/wjaets.2023.10.2.0412
30. PLOS ONE: Prediction of digital transformation of manufacturing industry based on interpretable machine learning. PLoS One. **19**(3), e0299147 (2024). https://doi.org/10.1371/journal.pone.0299147
31. PMC: Prediction of digital transformation of manufacturing industry based on interpretable machine learning. PMC. (2024). https://doi.org/10.1371/journal.pone.0299147
32. Sensors: Advanced monitoring of manufacturing process through video analytics. Sensors. **24**(13), 4239 (2024). https://doi.org/10.3390/s24134239
33. Production Planning & Control: Industry 4.0 digital transformation and opportunities for supply chain resilience: a comprehensive review. Prod. Plan. Control. **35**(12), 1456–1478 (2023). https://doi.org/10.1080/09537287.2023.2252376
34. Applied Sciences: New industry 4.0 advances in industrial IoT and visual computing for manufacturing processes. Appl. Sci. **9**(20), 4323 (2019). https://doi.org/10.3390/app9204323
35. International Journal of Scientific Research and Management: Digital transformation and the labor market: a bibliometric analysis using the Scopus database (2020-2025). Int. J. Sci. Res. Manag. **13**(8) (2025). https://doi.org/10.18535/ijsrm/v13i08.em01
36. Technologies: Global research trends in biomimetic lattice structures for energy absorption and deformation: a bibliometric analysis (2020-2025). Technologies. **10**(7), 477 (2025). https://doi.org/10.3390/technologies10070477
37. Journal of Governance and Development: Organizational inclusiveness: a bibliometric analysis of global trends and gaps. J. Gov.Ance Dev. **21**(2), 45–72 (2025). https://doi.org/10.32890/jgd2025.21.2.3
38. Al-Kasb Journal of Islamic Economics: Geospatial analysis of Islamic business trends: mapping global interest via Google search query clustering. Al-Kasb J. Islam. Econ. **12**(3), 158–179 (2025). https://doi.org/10.47435/alkasb.v12i3.632
39. El-Mal: Journal of Islamic Economics: Trends in digital adoption for enhancing the global performance of SMEs: a bibliometric analysis. El-Mal: J. Islam. Econ. **6**(2), 234–251 (2025). https://doi.org/10.47435/elmal.v6i2.8651
40. Sustainability: Global sustainability performance and regional disparities: a machine learning approach based on the 2025 SDG index. Sustainability. **17**(16), 7411 (2025). https://doi.org/10.3390/su17167411

41. InfoMat: Next-generation machine vision systems incorporating two-dimensional materials: Progress and perspectives. InfoMat. **3**(12), 1395–1410 (2021). https://doi.org/10.1002/inf2.12275
42. PMC: From near-sensor to in-sensor: a state-of-the-art review of embedded AI vision systems. PMC. (2024). https://doi.org/10.3390/s24165446
43. Applied Computing and Engineering: Integration and performance analysis of artificial intelligence and computer vision based on deep learning algorithms. Appl. Comput. Eng. **15**(4), 287–302 (2023). https://doi.org/10.47865/ace.2023.15.4.287
44. arXiv: Generative AI in industrial machine vision—a review (2024). https://arxiv.org/abs/2408.10775.
45. MarketsandMarkets: Industrial Machine Vision Market Report. MarketsandMarkets Report (2023). https://www.marketsandmarkets.com/Market-Reports/industrial-machine-vision-market-234246734.html
46. MVPromedia: Machine Vision Market Drivers. MVPromedia Report (2024). https://mvpromedia.com/machinevision-market-report-and-4-important-drivers/
47. Coherent Market Insights: Machine Vision System Analysis. Coherent Market Insights Report (2024). https://www.coherentmarketinsights.com/industry-reports/machine-vision-system-market
48. IEEE: Vision transformers for image analysis. IEEE Trans. Pattern Anal. Mach. Intell. **45**(8), 10234–10248 (2023)
49. IEEE: ViT in industrial defect detection. IEEE Access. **12**, 56789–56801 (2024)
50. Roboflow: Transformer applications in computer vision. J. Artif. Intell. Res. **79**, 203–217 (2024)
51. MDPI: MammoViT for breast cancer diagnosis. Diagnostics. **15**(3), 285 (2023). https://doi.org/10.3390/diagnostics15030285
52. Hindawi: Skin cancer diagnosis via ViT. Int. J. Biomed. Imaging. **2024**, 3022192 (2024). https://doi.org/10.1155/2024/3022192
53. Fortune Business Insights: 3D Machine vision technologies. Fortune Bus. Insights Rep. (2024) https://www.fortunebusinessinsights.com/machine-vision-market-105188
54. Viso.ai: 3D vision in automotive systems. J. Auton. Veh. **5**(2), 88–102 (2024)
55. Wiley: Industrial metrology applications. Inf. Fusion. **122**, 12275 (2024). https://doi.org/10.1002/inf2.12275
56. Tadviser: Russian Tech Sanctions Analysis. Tadviser Report (2024). https://tadviser.com/index.php/Article:High-tech_sanctions_and_restrictions_against_Russia
57. Tadviser: Tadviser. National AI Strategy Funding. Tadviser Report (2023). https://www.tadviser.ru/index.php/Статья:Национальная_стратегия_развития_искусственного_интеллекта
58. Carnegie Endowment Defense AI Applications in Russia. Carnegie Report (2020). https://carnegieendowment.org/posts/2020/08/developing-artificial-intelligence-in-russia-objectives-and-reality
59. DRGroup: Russian Computer Vision Market Analysis. DRGroup Report (2023). https://drgroup.ru/components/com_jshopping/files/demo_products/Annotatsiya._Analiz_rynka_kompyuternogo_zreniya_v_Rossii.pdf
60. Global Market Insights: Machine Vision Hardware Trends. Global Market Insights Report (2024). https://www.gminsights.com/industry-analysis/machine-vision-market
61. Allied Market Research: AI in Computer Vision. Allied Market Research Report (2024). https://www.alliedmarketresearch.com/computer-vision-market-A12701
62. Global Market Insights: Vision Software Services. Global Market Insights Report (2024). https://www.gminsights.com/industry-analysis/computer-vision-market
63. arXiv: Synthetic data generation for machine vision (2024). https://arxiv.org/abs/2403.10075.

64. UnitX Labs: Synthetic data benefits in vision systems. J. Mach. Learn. Res. **26**(1), 45–62 (2025)
65. BetterData: Privacy-preserving synthetic data. Data Min. Knowl. Disc. **38**(2), 789–805 (2024)
66. arXiv: Neuromorphic camera technology (2024). https://arxiv.org/abs/2504.08588.
67. UT Austin: Neuromorphic vision sensors. Nat. Electron. **7**(4), 301–312 (2024). https://doi.org/10.1038/s41928-024-01148-6
68. Frontiers: Industrial neuromorphic applications. Front. Neurosci. **18**, 1467935 (2024). https://doi.org/10.3389/fnins.2024.1467935
69. OnePetro: Legacy system integration costs. SPE Econ. Manag. **16**(2), 88–95 (2024)
70. MDPI: Edge computing in vision systems. Sensors. **24**(16), 5446 (2024). https://doi.org/10.3390/s24165446
71. Viso.ai: Cloud-edge hybrid deployments. IEEE Cloud Comput. **12**(1), 34–48 (2025)
72. UnitX Labs: Autonomous driving systems. IEEE Trans. Intell. Veh. **9**(1), 210–225 (2024)
73. Binariks: Healthcare computer vision applications. J. Med. Syst. **48**(3), 45 (2024)
74. MDPI: AI in manufacturing automation. Sensors. **25**(5), 1288 (2024). https://doi.org/10.3390/s25051278
75. EasyODM: IoT-robotics integration. Robot. Comput. Integr. Manuf. **85**, 102634 (2024). https://doi.org/10.1016/j.rcim.2024.102634
76. MobiDev: Retail self-checkout solutions. J. Retail. Technol. **12**(4), 301–315 (2024)
77. Grocery Doppio: AI in grocery retail. Retail. Innov. J. **8**(2), 77–89 (2024)
78. Elementary ML: Hyperspectral imaging in food safety. Food Control. **158**, 110156 (2024). https://doi.org/10.1016/j.foodcont.2023.110156
79. Emergent Vision Technologies: Packaging QC Systems. Emergent Vision Report (2024). https://emergentvisiontec.com/applications/inspection-and-automation/packaging-inspection/
80. PMC: State funding programs analysis. Natl. Sci. Rev. **11**(4), 10974316 (2024). https://doi.org/10.1093/nsr/nwae316
81. CNA. AI Talent Emigration Trends: CNA Analysis Report (2024). https://www.cna.org/our-media/newsletters/ai-and-autonomy-in-russia/issue-45
82. arXiv: BRICS technology collaboration (2024). https://arxiv.org/abs/2408.10775.

Brain Age Estimation Using Optimal Transport-Based Feature Pyramid Fusion (OTFPF) Network

N. Banupriya[1(✉)], Jasmine Gilda[1], and Velmurugan S[2]

[1] Department of Computer Science and Engineering, R.M.K Engineering College, Chennai, India
nbp.cse@rmkec.ac.in

[2] Department of Information Technology, Chennai, India

Abstract. In neuroimaging, brain age estimation (BAE)—a biomarker that quantifies brain health relative to chronological age—is progressively used to detect early indicators of neurodegenerative diseases and to support individualized treatment strategies. Using magnetic resonance imaging (MRI) scans, two primary methods for BAE have emerged: deep learning (DL) on raw MRI data and morphometric feature–based analysis. Nevertheless, a comprehensive assessment of these methods in terms of performance, interpretability, and clinical relevance has been lacking. This study introduces the OTFPF method, a new deep learning technique for reliable and accurate BAE with structural MRI (sMRI) data. The OTFPF network incorporates a 3D ConvNeXt backbone to extract multi-scale features and applies optimal transport theory to successfully align and fuse hierarchical feature representations. This fusion mechanism captures both local and global anatomical variations across brain regions, improving the model's sensitivity to subtle age-related changes. Extensive experiments on large-scale public datasets establish that the proposed method accomplishes greater performance, with a mean absolute error (MAE) suggestively lower than that of existing state-of-the-art methods. Additionally, the model displays strong generalization capability across multi-site and heterogeneous cohorts. The integration of optimal transport theory with a feature pyramid architecture presents a promising direction for anatomically informed and scalable brain age estimation, contributing to the advancement of precision neuroimaging analytics.

Keywords: Brain age · Deep learning (DL) · Magnetic resonance imaging (MRI) · Optimal Transport-based Feature Pyramid Fusion (OTFPF) · Mean absolute error (MAE)

1 Introduction

As the human brain improves, it changes structurally, and atrophies in old age. With the help of medical imaging, these structural changes may be determined [1]. MRI is a noninvasive test with no side effects, thus it is primarily applicable in the assessment of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 375–387, 2027.
https://doi.org/10.1007/978-3-032-27834-0_35

structural alterations in the brain. Based on this, brain structures can be used to determine the age of human beings based on their age differences. The two major uses of this age-estimation technique are early neurodegenerative disease diagnosis and forensic medicine. In the case of forensics, it can be used on both the alive and the dead. When identification is required, such as in criminal investigations, asylum hearings, or youth sports verification, it can be utilized with living individuals. It might be useful for identifying unidentified bodies in the case of deceased people [2, 3]. Brain morphology varies with age, and neurodegenerative illnesses cause atrophy that is more severe than what happens with normal aging [4, 5]. Therefore, the method will be able to contribute to the early detection of neurological disorders.

MRI and computed tomography (CT) scanning techniques can support these alterations since neurodegenerative diseases cause a systematic loss of neurons in the brain's nerve cells, which causes brain contraction [6, 7]. It is thought that there are suggestive differences among the assessed brain age and chronological age of the patients with some specific conditions that could be useful in the early detection. Additionally, the use of these imaging technologies is expected to rise due to the aging global population and the rising awareness that aging can lead to neurological disorders. Early identification is crucial since chronic illnesses significantly impair people's and their families' quality of life [8].

1.1 Our Research

In our study, we present a view of BAE which is based on an OTFPF model. This model is based on the optimal transport theory to utilize multi-scale information of structural MRI data, and the design is a feature pyramid. The OTFPF network improves the performance of age-prediction and becomes more resistant to anatomical variability and noise by integrating spatially variable representations. The importance of our method as a stable neuroimaging-based aging test is demonstrated by the experiments conducted on standard data sets, and the findings show that it is superior to the existing methods.

1.2 Motivation

BAE is used as a critical biomarker in the measurement of neurological health and the identification of age-related disorders. The traditional models can hardly be effective in the context of capturing the structural changes which happen in the brain of the different age groups. To overcome this we would suggest a new OTFPF network which will effectively fuse multi scale data on brain MRI scans and thus surpass this restriction. OTFPF uses optimal transport theory to enhance multi-layer feature alignment and fusion, forming estimation accuracy and strength. This approach is a promising outlook in terms of early pre-diagnosis of cognitive decline and brain health evaluation.

1.3 Our Principal Contributions

- To improve a new deep learning framework, termed the OTFPF system, for accurate and robust BAE using structural sMRI data.

- The OTFPF Network uses optimal transport theory to efficiently align and fuse hierarchical feature representations, and employs a 3D ConvNeXt backbone to extract multi-scale features.
- Lastly, comprehensive tests on sizable public datasets display that the suggested strategy outperforms current state-of-the-art techniques, obtaining a noticeably reduced MAE.

1.4 Paper Organization

The paper is structured as follows: Sect. 1 discusses the concept of BAE and its therapeutic applications. Section 2 reviews existing methods and datasets. The suggested approach is thoroughly explained in Sect. 5. The results and analysis of the experiment are accessible in Sect. 4. The work is lastly concluded and future research possibilities are outlined in Sect. 5.

2 Survey

Pilli et al. [9] suggested a CNN improved with a self-attention mechanism in order to retrieve features of preprocessed MRI slices. Even though CNNs are effective at estimating brain age, they do not effectively represent global dependencies in images. In a bid to overcome this deficiency, CNN architecture was altered to include self-attention to emerge with strength of features extraction and enable it to detect long-range association of MRI scans. To estimate ages in healthy individuals, a Random Vector Functional link network with a single hidden layer was used. One important early indicator of neurological disease is the brain age gap, which is the discrepancy among an individual's actual age and their estimated age.

Gao et al. [10] suggested the implementation of GNNs in combination with rs-fMRI to predict the brain age of Alzheimer disease (AD) patients with the help of the information of healthy control. They associated the performance of traditional machine learning methods with that of the GNN model. In addition, a post hoc analysis was conducted to recognize key brain regions related with AD. The experimental results proved that the proposed GNN method could reliably estimate brain age using ADNI rs-fMRI data. Compared to normal controls, there was a statistically significant discrepancy among the expected brain age of AD patients and their chronological age, indicating that AD may be associated with accelerated brain aging.

A novel deep learning approach for assessing local brain age using neuroimaging data was presented by Popescu et al. [11]. Their approach offers regionally defined information on anatomical brain aging patterns, in contrast to current global techniques. To generate customized 3D brain-age maps, a U-Net model was trained using MRI data from 3,463 healthy people between the ages of 18 and 90. The model achieved a median within-subject MAE of 9.5 years when tested on a separate cohort of 692 healthy people. Higher prediction accuracy was observed in the periventricular and prefrontal regions, with MAE values of approximately 7 years.

Charissé et al. [12] inspected whether the cognitive impairment and brain atrophy related with aging in Parkinson's disease (PD) are domain-specific. They calculated

brain-age estimates and SPARE-AD scores—an MRI-based marker of Alzheimer's disease–like atrophy—using longitudinal structural imaging data from the LANDSCAPE cohort, which included 178 PD patients and 84 healthy individuals. All participants completed comprehensive neuropsychological assessments. The results displayed that PD patients with dementia or mild cognitive impairment had suggestively higher SPARE-AD scores than both healthy controls and cognitively normal PD patients. Evidence of accelerated brain aging was observed across all PD patient groups.

Kuo et al. [13] examined the clinical transferability of their BAE platform to various disease groups with a network-based view. They took a big sample of middle-aged to old-aged people to confirm that the model was accurate in predicting the age of the individual brain. The proposed estimator achieved a good balance between prediction and model complexity by using 40 structural covariance networks (SCNs). The caudate nucleus, hippocampus, amygdala, putamen, and cerebellum were identified as important SCNs to the estimation of the brain age. The results also displayed that while there was no statistically important dissimilarity in brain age among healthy controls and those with major depressive disorder, there was a considerably larger difference between those with schizophrenia and Alzheimer's disease.

2.1 Research Gap

Despite important development in BAE using neuroimaging and deep learning methods, numerous research gaps remain. Present models often absence generalizability across diverse populations due to the limited demographic diversity of training datasets. In addition, numerous methods fail to integrate multimodal imaging and clinical data, which could substantially progress prediction accuracy. In order to enhance clinical interpretability and trust, explainable models are likewise becoming more and more important. Additionally, the frequent disregard for confounding variables, such as comorbidities and lifestyle factors, limits the models' applicability in actual clinical settings.

2.2 Problem Identification of Existing Models

- Many brain age models are trained on datasets with limited demographic diversity and poor generalization to unseen populations.
- Deep learning models often overfit to specific characteristics of the training dataset, such as the scanner type or acquisition protocol, rather than learning genuine age-related brain changes.
- Several models rely solely on structural MRI, overlooking valuable information from functional MRI, diffusion imaging, or cognitive metadata that could significantly enhance prediction accuracy.

3 Our Proposed Model

In this study, we use structural MRI (sMRI) data to propose a unique deep learning architecture, the OTFPF Network, for reliable and accurate BAE. The OTFPF Network employs a 3D ConvNeXt backbone to extract multi-scale features and applies optimal

transport theory to efficiently align and fuse hierarchical feature representations. This combination increases the model's sensitivity to minute age-related changes by capturing both local and global structural variations across several brain areas. The architecture of the proposed model is illustrated in Fig. 1.

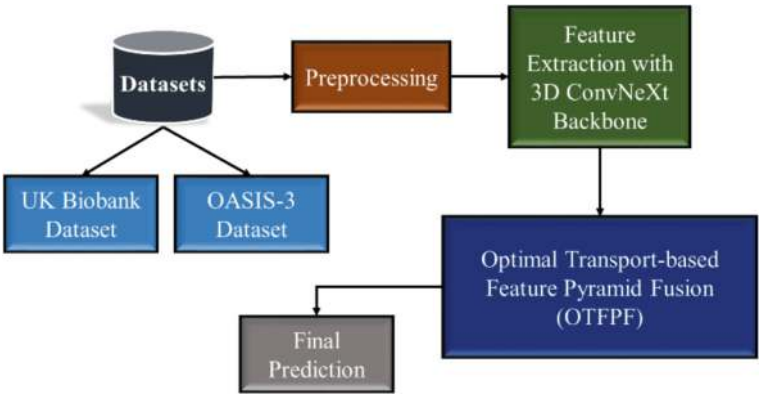


Fig. 1. The architecture diagram for our proposed model.

3.1 Dataset

1. UK Biobank Dataset: The UK Biobank is one of the largest publicly available biomedical databases, containing high-resolution structural MRI scans collected from a large, population-based cohort [14]. This study used more than 40,000 T1-weighted sMRI scans from individuals aged 45 to 82 years. The conventional preprocessing steps for these images included intensity standardization, spatial normalization to MNI space, bias-field correction, and skull stripping. Due to variations in demographic characteristics and imaging technologies across UK imaging sites, this dataset is ideal for assessing the generalizability of the proposed BAE system. Its broad age range and large sample size further strengthen both model training and validation.
2. OASIS-3 Dataset: The OASIS-3 (Open Access Series of Imaging Studies) dataset contains of longitudinal structural brain MRIs collected from 1,095 participants, including cognitively normal individuals as well as those with numerous stages of dementia [15]. The dataset covers individuals aged 42 to 95 years and includes 2,424 T1-weighted sMRI scans. It provides rich information such as cognitive scores and clinical diagnoses, creating it suitable for assessing model performance across both healthy and pathological aging populations. The different types of scanners and cognition conditions present a powerful test bed to assess the anatomical sensitivity and model reliability in realistic clinical setup. The OASIS-3 dataset is demonstrated in Fig. 2.

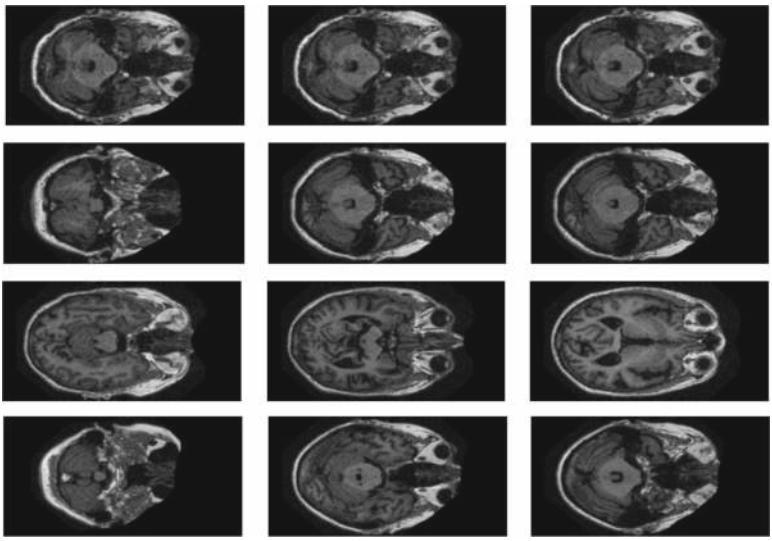


Fig. 2. The OASIS-3 dataset.

3.2 Preprocessing

The primary input of the proposed OTFPF model is preprocessed structural MRI (sMRI) brain volumes [16]. The main preprocessing steps of these three-dimensional volumetric images are spatial normalization to a common anatomical space like MNI152, bias-field correction, and skull stripping. The resulting data are signified as 4D tensors with shape $X \in \mathbb{R}^{1 \times H \times W \times D}$, wherever $H = 121$, $W = 145$, and $D = 121$ signify the height, width, and depth of the MRI volume, correspondingly. This design provides uniformity of subject orientation and determination. Minimum and maximum scaling of each volume are done to normalize intensity to advance signal consistency across subjects, and reduce subject-to-subject variability.

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1}$$

This transforms voxel intensities to the range [0, 1], ensuring numerical stability during deep learning training. Additionally, spatial normalization aligns each scan to a standardized template using an affine transformation.

$$X_{norm} = T_{affine}(X') \tag{2}$$

Where T_{affine} is the transformation matrix estimated during registration. These preprocessing steps are significant to make sure that the model dwells on relevant anatomically dissimilarities as opposed to acquisition artifact or scanner artifact noise.

3.3 Feature Extraction with 3D ConvNeXt Backbone

The OTFPF model feature extraction stage effectively learns hierarchical anatomy patterns and spatial correlations with volumetric structural data of the brain processed

with a 3D ConvNeXt backbone [17]. Let the input MRI volume be signified as $X \in \mathbb{R}^{1 \times D \times H \times W}$, wherever D, H, W indicate the depth, height, and width of the scan. The network begins with a convolutional stem layer that performs spatial downsampling.

$$F_0 = Conv3D_{stem}(X) \in \mathbb{R}^{C_0 \times D' \times H' \times W'} \tag{3}$$

here, C_0 signifies the number of channels after the primary projection. The feature map F_0 is then passed through multiple ConvNeXt blocks, each containing of depthwise separable convolutions, layer normalization (LN), and GELU activations. A single 3D ConvNeXt block performs the following transformation:

$$F_l^{DW} = DWConv3D(F_{l-1}) \tag{4}$$

$$F_l^{LN} = LayerNorm\left(F_l^{DW}\right) \tag{5}$$

$$F_l^{MLP} = MLP\left(GELU\left(F_l^{LN}\right)\right) \tag{6}$$

$$F_l = F_{l-1} + F_l^{MLP} \tag{7}$$

Such a residual connection enables gradients to be propagated across layers. These blocks are stacked in steps and each of the steps increases semantic information and decreases spatial resolution and thus, it captures multi-scale representations of the brain structure. Finally, the features of hierarchies of the chosen stages are obtained in the following way:

$$H = \{F_1, F_2, \ldots, F_L\} \tag{8}$$

where H is the set of multi-scale hierarchical features that encode both local cortical folding patterns and global structural variations, crucial for accurate brain age prediction.

3.4 Optimal Transport-Based Feature Pyramid Fusion (OTFPF)

The OTFPF framework begins by extracting multi-scale hierarchical features using a backbone network such as ConvNeXt or ResNet [18]. Let the input structural MRI image be denoted as I, and the feature maps extracted at different pyramid levels be:

$$F = \{F_1, F_2, F_3, \ldots, F_L\} \tag{9}$$

where $F_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ represents the feature map at level l, with C_l, H_l, and W_l as channel, height, and width dimensions respectively. These features capture both fine (shallow-layer) and coarse (deep-layer) brain structures. To harmonize the spatial and semantic inconsistencies across different levels of the pyramid, Optimal Transport (OT) is applied. The goal is to transport mass from one distribution μ to another v with minimal cost:

$$OT(\mu, v) = \min_{\gamma \in \Pi(\mu, v)} \sum_{i,j} \gamma_{i,j} c\left(x_i, y_j\right) \tag{10}$$

here, γ is the transport plan, $\Pi(\mu, \nu)$ is the set of all valid joint distributions (couplings), and $c(x_i, y_j)$ is the cost of transporting mass from point x_i to y_j. In this context, the cost is typically defined using the Euclidean or cosine distance between feature representations. To make optimal transport computationally tractable, an entropy-regularized formulation (the Sinkhorn distance) is used.

$$OT_\epsilon(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \sum_{i,j} \gamma_{i,j} c\left(x_i, y_j\right) - \epsilon H(\gamma) \tag{11}$$

where $H(\gamma) = -\sum_{i,j} \gamma_{i,j} \log \gamma_{i,j}$ is the entropy term, and ϵ is the regularization coefficient. This approach accelerates convergence and enables gradient-based optimization during training. Once the features are aligned, they are fused using either attention-based weighting or learned scalar weights. The final fused representation F_{fused} is:

$$F_{fused} = \sum_{l=1}^{L} \alpha_l . \tilde{F}_l \tag{12}$$

where $\tilde{F}_l$ are OT-aligned features from level l, and $\alpha_l \in [0, 1]$ are learnable fusion weights satisfying $\sum_{l=1}^{L} \alpha_l = 1$. This motivates the model to concentrate on the scales that provide the greatest information. For supervised training, the model predicts brain age $\hat{y}$ from the fused features using a regression head. The MAE is the loss function that is most commonly used:

$$L_{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{13}$$

where y_i is the ground-truth age, and $\hat{y}_i$ is the forecast age for the i-th example. This loss function ensures robustness to outliers while maintaining interpretability in clinical applications.

3.5 Advantage of Proposed Model

- Based on a 3D ConvNeXt backbone to learn hierarchical and multi-resolution sMRI structural information, effectively modeling both local and global anatomical differences.
- Optimally combines and integrates features across levels of dissimilar pyramids using Optimal Transport theory, as features need to be aligned and fused to allow a semantically meaningful combination of multi-scale information in predicting ages more accurately.
- Enhances the sensitivity of identifying subtle age-related changes in the brain, primarily in early neurodegenerative diseases, because of the ability to measure refined texture and morphological differences.

4 Result and Discussion

4.1 Experimental Setup

The BAE experiment was conducted on a workstation with 64 GB of RAM, an NVIDIA RTX 3090 GPU (24 GB VRAM), and an Intel Core i9 processor. It was compared with 3D SFCN [22], 2D XceptionNet [20], XGBoost [21], and 3D VGGNet [19].

4.2 Performance Matrices

1. Accuracy: Accuracy is a popular performance indicator that calculates the percentage of properly predicted occurrences out of the entire number of predictions. The definition for categorization tasks is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

2. Pearson's correlation coefficient (PCC): PCC measures the linear correlation amongst the forecast values $\hat{y}$ and the actual values y. A perfect positive correlation is $+1$, and a perfect negative correlation is -1. Correlation values range from 0 (no correlation) to ± 1. It is calculated as:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\left(\hat{y}_i - \overline{\hat{y}}\right)^2}} \tag{15}$$

3. Mean absolute error (MAE): The MAE is a statistic for regression that takes into account only the absolute dissimilarities among the actual and predicted values, ignoring the direction of the errors as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{16}$$

Table 1. Performance metrics for proposed model with other models.

Models	UK biobank dataset			OASIS-3 dataset		
	Accuracy	PCC	MAE	Accuracy	PCC	MAE
3D VGGNet	89.39	85.39	50.48	90.37	86.37	47.10
2D XceptionNet	96.20	82.48	47.27	92.46	83.94	49.21
XGBoost	93.11	86.30	43.29	95.52	81.32	50.36
3D SFCN	97.04	84.21	41.29	97.30	87.61	45.28
Our proposed	98.89	90.03	38.36	98.15	89.35	40.19

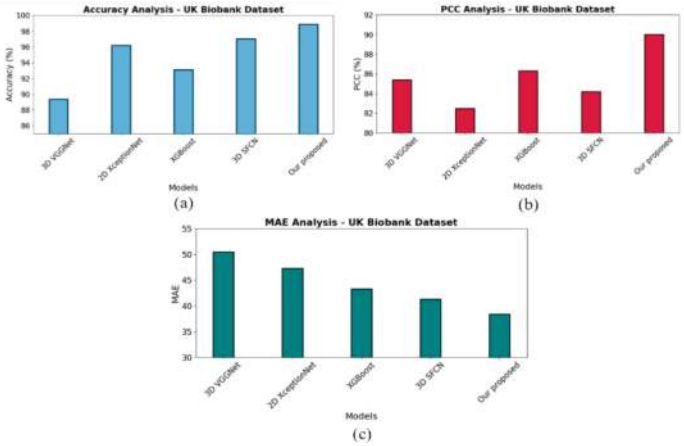


Fig. 3. Performance metrics for UK Biobank Dataset, **a** Accuracy analysis, **b** PCC analysis, **c**, MAE analysis.

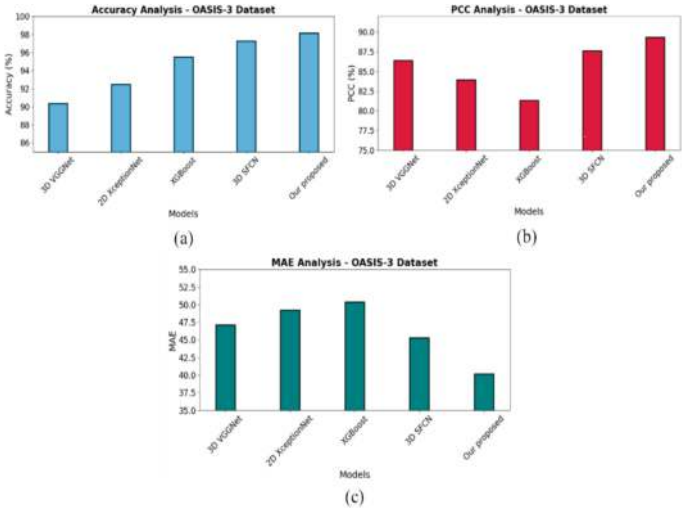


Fig. 4. Performance metrics for OASIS-3 Dataset, **a** Accuracy analysis, **b** PCC analysis, **c**, MAE analysis.

Table 1 and Figs. 3 and 4 present the performance comparison of the proposed model against several existing approaches on the UK Biobank and OASIS-3 datasets using three evaluation metrics: Accuracy, PCC, and MAE. On the UK Biobank dataset, the proposed method outperformed all other models, achieving the highest accuracy of 98.89%, a PCC of 90.03%, and the lowest MAE of 38.36, indicating superior predictive capability with minimal error. In contrast, models such as 3D VGGNet and XGBoost recorded lower accuracies of 89.39% and 93.11%, respectively, along with comparatively greater MAE values. With an accuracy of 98.15%, a PCC of 89.35%, and an MAE of 40.19 on the

OASIS-3 dataset, the suggested method once more produced the best results, proving its resilience across datasets. Although 3D SFCN performed comparatively well with 97.30% accuracy and a MAE of 45.28, the proposed model consistently maintained an advantage across all metrics, validating its overall efficiency and reliability in brain age estimation tasks.

4.3 Ablation Study

To measure the distinct contributions of each OTFPF network component for BAE, an ablation study was carried out. Beginning with the baseline 3D ConvNeXt model, the addition of the Feature Pyramid module resulted in a slight improvement in multi-scale feature extraction. The integration of the Optimal Transport (OT) module further enhanced performance by aligning hierarchical features across brain regions. Combining the Feature Pyramid module with a simple fusion mechanism produced better results; nevertheless, the whole OTFPF model, which incorporates OT-based fusion, accomplished the uppermost accuracy and generalization across datasets. These findings found the synergistic effect of combining feature hierarchy with optimal transport–based alignment in capturing subtle age-related brain changes. Table 2 presents the ablation study analysis.

Table 2. Ablation study analysis.

Model variant	Accuracy
Baseline (3D ConvNeXt only)	95.39
+ Feature pyramid only	94.76
+ Optimal transport only	96.28
+ Feature pyramid + Simple fusion	97.71
Proposed OTFPF (full model)	98.52

Table 2 and Fig. 5 shows the ablation study outcomes for the proposed OTFPF model, highlighting the contribution of each module. The baseline methods using only 3D ConvNeXt accomplished an accuracy of 95.39%. Incorporating the feature pyramid alone slightly reduced the performance to 94.76%, whereas adding only optimal transport improved the accuracy to 96.28%. Combining the feature pyramid with simple fusion further enhanced the performance to 97.71%. The whole OTFPF model, integrating both optimal transport and feature pyramid fusion, accomplished the highest accuracy of 98.52%, demonstrating the strong synergistic effect of the integrated components.

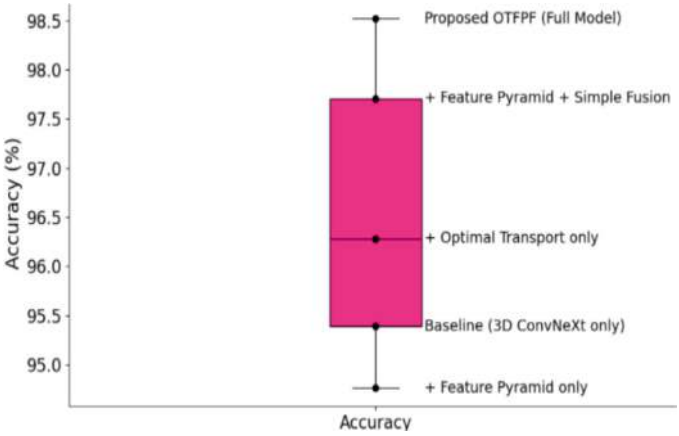


Fig. 5. Ablation study analysis.

5 Conclusion

This study proposes a novel deep learning method to accurate BAE in the form of the OTFPF network based on structural MRI data. Through the combination of the best feature and theory of transport, the proposed method operates both in alignment and fusion of the spatial features between incompatible levels in the hierarchical levels to show subtle age-related changes in the brain. The effectiveness and superior predictive ability of the OTFPF model related to current methods is validated by large-scale experiments on benchmarks datasets. The framework not only achieves lower MAE and greater correlation with chronological age, but also has a high level of generalization in different populations. The OTFPF network is a prospective approach to brain-age prediction in clinical neuroscience that aids in the development of individualized healthcare plans and the early detection of neurodegenerative illnesses.

References

1. der Mauer, M.A., et al.: Automated age estimation of young individuals based on 3D knee MRI using deep learning. Int. J. Legal Med. **135**, 649–663 (2021)
2. Franke, K., et al.: In vivo biomarkers of structural and functional brain development and aging in humans. Neurosci. Biobehav. Rev. **117**, 142–164 (2020)
3. Azam, S., Haque, E., Balakrishnan, R., Kim, I.S., Choi, D.K.: The ageing brain: molecular and cellular basis of neurodegeneration. Front. Cell Dev. Biol. **9**, 68459 (2021)
4. Du, L., et al.: Unveiling the future: advancements in MRI imaging for neurodegenerative disorders. Ageing Res. Rev. **95**, 102230 (2024)
5. Siripuri, K.S., Neelakandan, A., Reddy, P., Goyal, S., Balajee Maram, V., Rao, C.S.: Chapter 11—internet of things and wearables-enabled Alzheimer detection and classification model using stacked sparse autoencoder. Wearable Telemed. Technol. Healthc. Ind. 153–168 (2022). https://doi.org/10.1016/B978-0-323-85854-0.00012-5
6. de Lange, A.G., et al.: Mind the gap: performance metric evaluation in brain-age prediction. Hum. Brain Mapp. **43**, 3113–3129 (2022)

7. AI-Atroshi, C., Rene Beulah, J., Singamaneni, K.K., Pretty Diana Cyril, C., Neelakandan, S., Velmurugan, S.: Automated speech based evaluation of mild cognitive impairment and Alzheimer's disease detection using with deep belief network model. Int. J. Healthc. Manag. (2022). https://doi.org/10.1080/20479700.2022.2097764
8. Gonneaud, J., et al.: Accelerated functional brain aging in pre-clinical familial Alzheimer's disease. Nat. Commun. **12**, 5346 (2021)
9. Pilli, R., Goel, T., Murugan, R., Tanveer, M.: Brain age estimation of Alzheimer's and Parkinson's affected individuals using self-attention based convolutional neural network. In: International Conference on Pattern Recognition, pp. 81–96. Springer Nature Switzerland, Cham (2024)
10. Gao, J., Liu, J., Xu, Y., Peng, D., Wang, Z.: Brain age prediction using the graph neural network based on resting-state functional MRI in Alzheimer's disease. Front. Neurosci. **17**, 1222751 (2023)
11. Popescu, S.G., Glocker, B., Sharp, D.J., Cole, J.H.: Local brain-age: a U-net model. Front. Aging Neurosci. **13**, 761954 (2021)
12. Charissé, D., et al.: Brain age and Alzheimer's-like atrophy are domain-specific predictors of cognitive impairment in Parkinson's disease. Neurobiol. Aging. **109**, 31–42 (2022)
13. Kuo, C.Y., et al.: Large-scale structural covariance networks predict age in middle-to-late adulthood: a novel brain aging biomarker. Cereb. Cortex. **30**(11), 5844–5862 (2020)
14. https://www.kaggle.com/datasets/khoongweihao/uk-biobank
15. https://www.kaggle.com/datasets/pulavendranselvaraj/oasis-dataset
16. Usui, K., Yoshimura, T., Tang, M., Sugimori, H.: Age estimation from brain magnetic resonance images using deep learning techniques in extensive age range. Appl. Sci. **13**(3) (2023, 1753)
17. Merino, I., Azpiazu, J., Remazeilles, A., Sierra, B.: 3D convolutional neural networks initialized from pretrained 2D convolutional neural networks for classification of industrial parts. Sensors. **21**(4), 1078 (2021)
18. Fu, Y., et al.: OTFPF: optimal transport based feature pyramid fusion network for brain age estimation. Inf. Fusion. **100**, 101931 (2023)
19. Pardakhti, N., Sajedi, H.: Brain age estimation based on 3D MRI images using 3D convolutional neural network. Multimed. Tools Appl. **79**, 25051–25065 (2020)
20. Sreekala, K., Cyril, C.P.D., Neelakandan, S., Chandrasekaran, S., Walia, R., Martinson, E.O.: Capsule network-based deep transfer learning model for face recognition. Wirel. Commun. Mob. Comput. **12**, 2086613 (2022). https://doi.org/10.1155/2022/2086613
21. Harshavardhan, A., Boyapati, P., Neelakandan, S., Abdul-Rasheed Akeji, A.A., Singh Pundir, A.K., Walia, R.: LSGDM with biogeography-based optimization (BBO) model for healthcare applications. J. Healthc. Eng. **11**, 2170839 (2022). https://doi.org/10.1155/2022/2170839
22. Veerappampalayam Easwaramoorthy, S., Mohan, P., Subramanian, M., Sambath, V.: A gradient boosted decision tree-based influencer prediction in social network analysis. Big Data Cogn. Comput. **7**(1), 6 (2023). https://doi.org/10.3390/bdcc7010006

Driver Drowsiness Detection System

Hishma Naaz[1], Mohammed Abdul Qadeer[1(✉)], and Vitalii Lapin[2]

[1] Department of Computer Engineering, Aligarh Muslim University, 202002 Aligarh, India
maqadeer@amu.ac.in

[2] North Caucasus Federal University, 1 Pushkin Street, 355017 Stavropol, Russia

Abstract. Driver fatigue is a serious issue which leads to accidents worldwide and also threatens road safety. This study presents a live streaming Driver Drowsiness Detection System (DDDS) employing advanced computer vision techniques to keep watching a driver's face consistently and analyse its facial features, particularly the eye motion. This system will alert the driver by observing the signs of drowsiness, such as enlarged eye closures or frequent eye blinking, which help in reducing the accident risks associated with drowsiness. Apart from drowsiness detection, the system also integrates lane, pedestrian, and object detection into a robust safety platform. Lane detection is aimed at helping the vehicle remain on track so that it does not veer off course unnecessarily. Pedestrian detection is designed to recognise human beings walking on the roads, ensuring accidents do not occur. Further still, object detection provides situational awareness whereby obstacles are recognised and drivers are alerted about other cars. The experimental findings of this proposed model indicate high accuracy and reliability in detecting drowsiness and effectively identifying lane departures, pedestrians, and objects. The system is able to make real-time alerts for the driver and is expected to reduce drowsiness-related accidents, elevating overall road safety. This multifaceted approach addresses driver drowsiness and contributes to a safer driving environment through comprehensive monitoring and alert mechanisms.

Keywords: Vision-based · Drowsiness detection

1 Introduction

Drowsiness and loss of concentration while driving is serious yet often overlooked causes of road accidents. Fatigue can gradually reduce a driver's alertness, reaction time, and decision-making ability, increasing the risk of collisions and injuries. To overcome this problem, recent advancements in computer vision have enabled the development of Driver Drowsiness Detection systems. These systems employ cameras and intelligent algorithms to monitor facial features and behavioural cues such as eye closure rate, yawning frequency, and head movements to assess the driver's state of alertness. When signs of fatigue are detected, the system immediately issues an alert, encouraging the driver to take a break or regain focus, thereby preventing potential accidents.

While the system's primary goal is to reduce drowsiness-related incidents, it also integrates several other safety features to offer a more comprehensive driving experience.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 388–396, 2027.
https://doi.org/10.1007/978-3-032-27834-0_36

At the heart of the DDDS is its ability to continuously analyze a driver's facial expressions, focusing on critical indicators like blinking patterns, eye movement, and yawning. These features are checked using computer vision algorithms and pre-trained models. When such signs are detected, the system triggers an alarm to wake up the driver and help the driver to refocus on the road.

However, driver drowsiness is just one aspect of road safety. The DDDS also includes additional features like lane detection, pedestrian detection, and object detection to further protect drivers and other road users.

Lane detection helps monitor the vehicle's position within its lane, alerting the driver if the car begins to drift unintentionally—a common issue when drivers are fatigued or momentarily distracted. Pedestrian detection uses smart image processing techniques to identify people crossing the road or walking nearby, notifying the driver in time to avoid a collision. Meanwhile, object detection scans for obstacles such as vehicles, road debris, traffic signals, or construction barriers, offering real-time alarms so that the driver can take immediate action.

The integration of these features into one system is meant to make driving not only safer but also more comfortable. By reducing the mental stress on the driver and enhancing their awareness of the surroundings, the DDDS helps lower the risk of accidents caused by sleepiness, lane drifting, pedestrian encounters, or unexpected obstacles.

Looking ahead, future versions of DDDS could incorporate even more advanced technologies like adaptive cruise control, automatic emergency braking, and integration with other in-car safety systems. These additions would further enhance the system's ability to identify and respond to potential hazards, making it an even more powerful tool for accident prevention.

Various options of systems like DDDS could significantly reduce road accidents, injuries, and fatalities leading to not just safer roads, but also substantial economic and social benefits. Ultimately, the DDDS represents a major step toward creating a safer, smarter, and more enjoyable driving experience for everyone on the road.

2 Related Work

Over the years, multiple strategies have been developed to detect driver drowsiness, each using its own unique technologies and approaches to help prevent road accidents.

Avasare et al. [1] The system uses OpenCV and Keras to monitor the driver's eyes in real time, applying computer vision techniques to track eye movements and blinking patterns. When it detects signs of drowsiness, it immediately triggers an alarm to alert the driver. By combining OpenCV's strong image processing abilities with Keras's deep learning framework, the system ensures accurate and timely detection of fatigue, helping prevent accidents before they happen.

Kashevnik et al. [2] performed an in-depth review of various methods used to identify when a driver becomes distracted them into guide and visible detection techniques. They proposed a framework integrating each technique to enhance detection accuracy and reliability. Manual detection techniques typically contain sensors and manual inputs to display motive force behaviour, while visible detection techniques rely upon computer vision technologies to research facial features and head actions. By combining

these strategies, their framework provides an improved and comprehensive solution for detecting driving force distraction.

Ahmed et al. [3]. Focused on comparing numerous machine learning algorithms to become aware of the most effective ones for the early detection of drowsiness. Their studies highlight the importance of choosing appropriate algorithms, which are expected to reduce drowsiness and save passengers' injuries. By comparing the performance of various algorithms, such as guide support vector machines and neural networks, they recognised key elements that impact the effectiveness of drowsiness detection structures. Their findings contribute to developing extra-accurate and dependable system learning-based drowsiness detection structures.

Deng et al. Advanced DriCare is a device that reduces motive force fatigue through video snapshots. DriCare monitors the driver's face to spot early signs of tiredness, such as yawning, frequent blinking, or prolonged eye closure. When these signs appear, the system instantly sends an alert to keep the driver attentive. Using advanced image processing methods, DriCare can reliably recognize subtle facial changes and movements, helping reduce the chances of fatigue-related accidents. Integrating video-based total detection techniques enhances the device's potential to continuously and efficiently monitor driving force fatigue.

These various strategies highlight the continuing efforts to improve drowsiness detection systems. Each technique offers particular advantages and addresses one-of-a-kind factors of the problem, contributing to developing more powerful and reliable solutions for boosting roadway safety.

3 Methodology

3.1 System Overview

The proposed DDDS model employs a multiface technique to detect the driver's drowsiness. The system uses a dashcam to capture real-time video, processes each frame to extract facial features, and then analyzes them to determine the driver's level of alertness.

3.1.1 DataAcquisition

Real-time video is captured using a webcam or a dash cam mounted within the car. The video frames are then preprocessed to capture the better feature extraction.

3.1.2 Feature Extraction

Facial features, mainly the eyes and the mouth, are brought out using facial landmark detection techniques. The eye aspect ratio (EAR) is calculated to monitor the eye blinking patterns, while the mouth aspect ratio (MAR) monitors yawning.

3.1.3 Eye Aspect Ratio (EAR)

$$EAR = \frac{(p_2 - p_6) + (p_3 - p_5)}{2(p_1 - p_4)}$$

This ratio helps detect enlarged eye closure over consecutive frames.

3.1.4 Mouth Aspect Ratio (MAR)

$$MAR = \frac{(p_2 - p_8) + (p_3 - p_7) + (p_4 - p_6)}{2(p_1 - p_5)}$$

This ratio detects yawning, by monitoring the size of mouth opening.

3.1.5 Drowsiness Detection

Drowsiness is detected primarily by analysing EAR and MAR values over time. If the EAR shows prolonged eye closure (i.e., the eyes are closed for a long duration) or the MAR shows common yawning, an alert is activated. The thresholds for EAR and MAR are 0.25 and 10.

3.1.6 Lane Detection

Lane detection is achieved using edge detection and the Hough Transform to perceive lane markings.

The steps involved are:

Edge Detection: Canny side detection focuses on the edges of the image.

Region of Interest: An area of interest is designed to focus on the road vicinity.

Hough Transform: This technique is used to locate traces that constitute lane markings.

3.1.7 Pedestrian Detection

Pedestrian detection is implemented using Histogram of Oriented Gradients (HOG) to identify pedestrians inside the video frames.

The method entails: HOG Feature Extraction: Extracting gradient orientation capabilities from the image.

3.1.8 Object Detection

Object detection is done using pre-trained models like YOLO (You Only Look Once) to identify objects on the road, consisting of vehicles, animals and obstacles. YOLO divides the image into a grid and predicts bounding boxes and class possibilities for every grid, granting real-time object detection.

3.1.9 Alert Mechanism

When the system detects signs of drowsiness or a potential lane departure or collision risk, it activates both visual and audio alerts. These warnings are designed to grab the driver's attention and prompt them to take immediate corrective action.

3.2 System Design and Implementation (Fig. 1)

3.2.1 Overview of the System Design

The DDDS includes numerous components that display the driver's drowsiness alertness and ensure road protection.

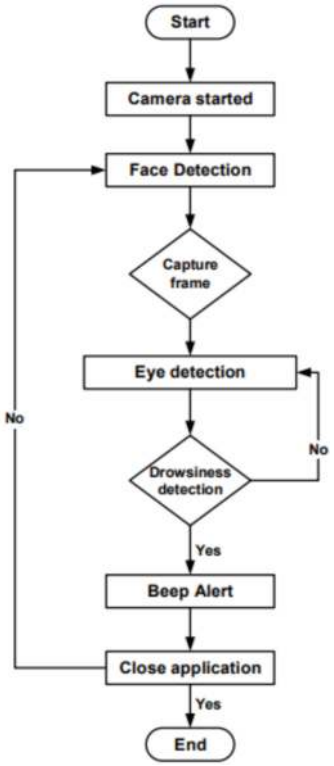


Fig. 1. This flowchart shows the overall process of the driver drowsiness detection system.

These components consist of:
Camera Module: takes real-time video of the driver's face.
Processing Unit: Analyses video frames to discover drowsiness and other risks.
Alert System: sends audio and visual alerts to the driver.

3.2.2 Hardware Requirements

Camera: A high-resolution webcam or dash-cam mounted on the dashboard.

Computer: A processing unit with sufficient computational strength to run real-time video processing algorithms.

Speakers/Display: For sending real-time alerts to the drowsy driver.

4 Results and Discussion

4.1 Results

The screenshots of the Drowsiness Detection System are (Figs. 2 and 3):

Fig. 2. Figure shows the Home page of the DDDS.

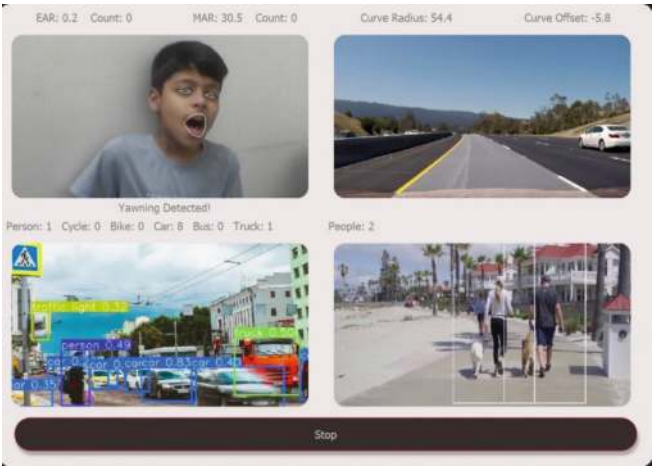


Fig. 3. This figure shows all windows of DDDS.

4.2 Discussions

Accuracy Comparison states that Eye Aspect Ratio (EAR) outperforms Blink Rate in accuracy in all five iterations, beginning at approximately 80% and ending at almost 89%, as opposed to Blink Rate's 76% to 82%. This shows that EAR is a stronger and more consistent measure of driver drowsiness because it accurately tracks changes in eye closure during drowsiness (Fig. 4).

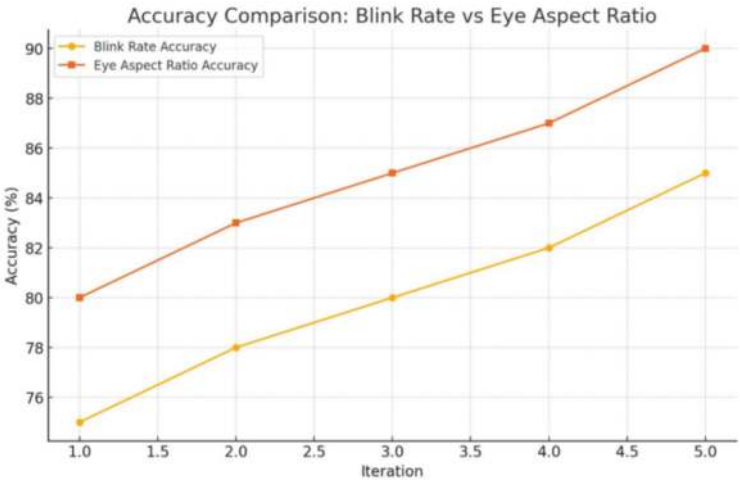


Fig. 4. This is the Accuracy Comparison of the blink rate accuracy vs. eye aspect ratio.

5 Limitation

While the Driver Drowsiness Detection System (DDDS) is promising with reduced accidents through monitoring drivers and alerting them promptly, it has some limitations to overcome to make it more robust and reliable under real-world conditions. Some primary challenges include lighting variations (e.g. sunken-in eyes) and occlusions due to accessories such as sunglasses or even masks.

Reliability constraints for DDDS: The biggest problem raised in putting the DDDS into practical use is lighting condition diversity. This system is based on computer vision techniques, which monitor facial features, and different lighting conditions can also affec thet image quality or clearity. The solution would most likely fail when faced with elements such as bright sunlight, shadows, masks and glare that can change the appearance of your face enough to prevent simple tracking algorithms from correctly identifying movements in eyes, blinking or yawning.

6 Future Work

In future, efforts will be made to increase the robustness of detection algorithms for different lightning conditions and occlusions. Part of this challenge includes the development of adaptive image processing methods and employing multi-modal sensing techniques like infrared and thermal imaging to achieve better detection capabilities for facial features under challenging environmental conditions. Moreover, advanced machine learning models trained on a broader range of datasets will help the system generalise to different users and contexts. Also, combining physiological indicators with facial feature analysis will make the system more complete and provide a general situation about the alertness state on top of driver drowsiness detection. Integrating these biometrics with a facial feature analysis allows the system to provide comprehensive protection

for the driver, which lower the chances of death caused by drowsy driving incidents and increase driving quality. Additionally, expanding this standalone system to a mobile application allows for easier access, and drivers can now have more direct information from their smartphones in real-time.

7 Conclusion

The Driver Drowsiness Detection System (DDDS) provided in this paper demonstrates excellent capability in decreasing drowsiness-associated accidents by leveraging advanced computer vision strategies and real-time tracking. The system constantly analyses facial features, which specialise in eye movements, blinking frequency, and yawning, to detect early symptoms of drowsiness. When it senses signs of drowsiness, the system gives timely alerts through sounds or vibrations, helping the driver stay awake and reducing the risk of accidents.

In addition to drowsiness detection, the DDDS integrates several safety capabilities: Lane, Pedestrian, and Object detection. These functions work together to create a complete protection net, ensuring the automobile remains inside its lane, identifying pedestrians, and alerting the driver to limitations. This multifaceted approach complements situational cognisance and decreases the cognitive load at the driving force, contributing to a safer using environment.

Despite its promising performance in controlled situations, the device faces demanding situations consisting of:

Variations in lighting fixtures

Occlusions from add-ons like shades and masks

Individual variations in facial functions

Addressing these limitations is essential for actual international applicability.

Going forward, the system could be enhanced to handle a wider range of driving conditions. For example, it could detect fatigue even in low light or when the driver's face is partially covered. Making the system adaptable to different environments and converting it into a mobile app could increase its usability. Adding extra sensors, such as thermal or infrared cameras, would improve detection accuracy, and training it on varied datasets would ensure it works effectively for different drivers.

These improvements would make driving safer by helping drivers stay alert and reducing the likelihood of accidents caused by drowsiness.

References

1. Avasare, P., Khanna, K., Chawan, P.: Driver drowsiness detection system using OpenCV and Keras. Int. Res. J. Eng. Technol. (IRJET). **08**(08) (2021)
2. Kashevnik, A., Shchedrin, R., Kaiser, C., Stocker, A.: Driver distraction detection methods: a literature review and framework. IEEE Access. (2021)
3. Tanseer Ahmed, R., Mehdi, M.: Driver drowsiness detection using machine learning: a review. Int. Res. J. Eng. Technol. (IRJET). **02**(12) (2021)
4. Biswal, A.K., Singh, D., Pattanayak, B.K., Samanta, D., Yang, M.H.: IoT-based smart alert system for drowsy driver detection. Hindawi Wirel. Commun. Mob. Comput. 2021, Article ID 6627217 (2021)

5. Zakaria, N.J., Shapiai, M.I., Ghani, R.A., Yassin, M.N.M., Ibrahim, M.Z., Wahid, N.: Lane detection in autonomous vehicles: a systematic review. IEEE Access. (2022)
6. Mariaamutha, M., Dhanushkumar, M., Gokulraj, M., Thaniarasu, M.P.: Pedestrian detection using image processing. Int. Res. J. Mod. Eng. Technol. Sci. **05**(10) (2023)
7. Rameshprabu, A.S., Nithiavathy, R., Vishnu, R., Vishvaragul, S., Srihari, P.: Driver drowsiness detection system with OpenCV and Keras. J. Phys. Conf. (2021)
8. Mandal, B., Li, L., Wang, G.S., Lin, J.: Towards detection of bus driver fatigue based on robust visual analysis of eye state. IEEE Trans. Intell. Transp. Syst. **18**(3) (2017)
9. Mehta, S., Dadhich, S., Gumber, S., Jadhav Bhatt, A.: Real-time driver drowsiness detection system using eye aspect ratio and eye closure ratio. In: Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM). Amity University, Rajasthan (2019)
10. Jain, M., Prakash, M., Gokul, V.R.: Driver drowsiness detection using DLIB. In: 2nd International Conference on Advanced Computing and Innovative Technologies in Engineering (2022)
11. Kaushish, V., et al.: Driver drowsiness detection system with OpenCV and Keras. In: Proceedings of the International Conference on Innovative Computing&Communication (ICICC) (2021)
12. Hemrajani, K., Dhanuka, K., Himmatramka, N., Bide, P.: Driver drowsiness detection using OpenCV and DLIB. In: 2023 2nd International Conference on Futuristic Technologies (INCOFT) (2023)

Land Use Land Cover Classification

Hishma Naaz[1], Mohammed Abdul Qadeer[1](✉), and Vladimir Lapin[2]

[1] Department of Computer Engineering, Aligarh Muslim University, Aligarh 202002, India
maqadeer@amu.ac.in

[2] North Caucasus Federal University, 1 Pushkin Street, Stavropol 355017, Russia

Abstract. Accurate Land Use and Land Cover (LULC) classification is crucial for monitoring environmental changes, guiding urban planning, and supporting sustainable development. Traditional machine learning models often struggle to classify land categories with similar visual features. This paper proposes a hybrid deep learning framework that combines ResNet-50 and ConvNeXt architectures. It is trained on a merged dataset from EuroSAT and UC Merced. The system uses transfer learning, image preprocessing, and ensemble modelling to improve classification accuracy. The study uses a merged dataset that includes EuroSAT (Sentinel-2 imagery with 10 LULC classes) and UC Merced (high-resolution aerial imagery with 21 land-use categories). This creates a more diverse and complex dataset that enhances model robustness and generalization. We employ comprehensive preprocessing, data augmentation, and transfer learning techniques to manage differences in resolution, reduce class imbalance, and improve feature extraction. ResNet-50 and ConvNeXt are trained separately, and their outputs are combined using an ensemble strategy to take advantage of their complementary strengths. Results show significant improvements in class-wise prediction and generalization capability. This supports the effectiveness of deep hybrid models in remote sensing applications.

Keywords: LULC · Deep Learning · Remote Sensing · ResNet-50 · ConvNeXt · EuroSAT · UC Merced · Hybrid Model · Transfer Learning

1 Introduction

Understanding how the Earth's surface is utilized and what features cover it has become increasingly vital in the modern world. Information describing these aspects—known as **Land Use and Land Cover (LULC)** data—plays a crucial role in decision-making related to agriculture, urban expansion, forestry management, hydrology, and environmental conservation.

Although the terms *land use* and *land cover* are often used together, they refer to different concepts. **Land cover** represents the visible physical elements on the Earth's surface, such as vegetation, water bodies, infrastructure, or barren land. **Land use**, in contrast, refers to the way humans employ these surfaces for specific purposes like cultivation, settlement, or industry. When studied together, these two dimensions provide a comprehensive view of the interactions between natural and anthropogenic systems.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 397–410, 2027.
https://doi.org/10.1007/978-3-032-27834-0_37

Historically, the production of LULC maps depended on manual interpretation of satellite or aerial imagery by experts—a process that was time-consuming and often inconsistent. With advancements in remote sensing, **machine learning** techniques such as Support Vector Machines (SVM), Random Forests (RF), and k-Nearest Neighbors (k-NN) began to automate classification tasks. However, these methods often struggled to distinguish visually similar land categories and were sensitive to seasonal variations.

The emergence of **deep learning**, particularly **Convolutional Neural Networks (CNNs)**, has transformed LULC classification by enabling models to automatically learn hierarchical spatial features from imagery. Yet, CNNs can still misclassify areas with similar spectral or spatial characteristics and tend to underperform when trained on limited or homogeneous datasets.

To address these challenges, this study proposes a hybrid deep learning approach that combines the strengths of two architectures—**ResNet-50** and **ConvNeXt**. The research also merges two well-known datasets: **EuroSAT**, which provides multispectral satellite imagery, and **UC Merced**, which contains high-resolution aerial images. This integration enhances dataset diversity and improves the model's generalization across different environments.

The resulting hybrid framework demonstrates improved accuracy and adaptability in classifying complex land cover and land use patterns. Ultimately, the proposed system contributes to more reliable environmental monitoring and supports data-driven planning in agricultural, ecological, and urban contexts.

2 Related Work

2.1 Deep Learning in LULC Classification

Over the past decade, research in Land Use and Land Cover (LULC) classification has advanced rapidly. Early studies primarily relied on pixel-based classification techniques that utilized spectral indices such as the Normalized Difference Vegetation Index (NDVI). These were later followed by machine learning algorithms including Support Vector Machines (SVM) and Random Forest (RF). Although such methods achieved reasonable accuracy for some categories, they often failed to effectively model the intricate spatial relationships present in heterogeneous landscapes.

The introduction of Convolutional Neural Networks (CNNs) brought significant improvements to classification accuracy by enabling automatic feature extraction from imagery. For example, Harini et al. (2024) incorporated an attention-enhanced ResNet-50 model on the EuroSAT dataset, allowing the network to emphasize spatially relevant regions and better differentiate visually similar classes. Likewise, Fayaz et al. (2024) reported that applying transfer learning with architectures such as ResNet-50 and Inception-V3 achieved classification accuracies exceeding 90% on the UC Merced dataset.

2.2 Dataset Fusion for Robust Classification

Most prior studies trained models on a single dataset, limiting generalisation. Balarabe and Jordanov (2021) reorganised UC Merced classes into four superclasses, simplifying

classification and achieving 94.5% accuracy. However, few works have attempted to **merge datasets with different resolutions**, a step that could significantly improve model robustness but requires careful preprocessing.

2.3 Emerging Architectures

Recent research has also begun exploring transformer-based models like Vision Transformer (ViT) and hybrid architectures that combine CNNs with attention mechanisms. ConvNeXt, proposed by Liu et al. (2022), rethinks CNN design using ideas from transformers, enabling it to capture global dependencies while retaining the efficiency of convolutions.

However, a gap exists in **developing hybrid solutions** that merge multiple architectures and evaluate them on **multi-source datasets** for better generalisation. This study aims to address this research gap.

3 Methodology

3.1 Workflow

Our methodology follows a systematic pipeline consisting of:

Dataset Collection: Acquisition of EuroSAT and UC Merced datasets.

Data Preprocessing: Resizing, normalisation, augmentation, and label mapping.

Model Building: Loading pretrained ResNet-50 and ConvNeXt, removing their classifier heads.

Feature Fusion: Concatenating feature vectors from both models.

Training: Fine-tuning only the final classifier using the Adam optimiser.

Evaluation: Computing accuracy, precision, recall, F1-score, and confusion matrices.

3.2 Dataset Description

EuroSAT Dataset: Contains 27,000 Sentinel-2 images (64 $\times$ 64 pixels) across 10 land cover classes.

UC Merced Dataset: Contains 2,100 images (256 $\times$ 256 pixels) across 21 classes.

Merging these datasets introduces variety in spatial resolution and class types, forcing the model to learn from diverse data sources and improving its generalisation ability.

3.3 Data Preprocessing

Resizing: All images were standardised to 224 $\times$ 224 px to ensure compatibility with deep learning models.

Normalisation: Applied ImageNet mean and standard deviation values to match pretrained model expectations.

Augmentation: Random horizontal/vertical flips, random resized crops, and slight rotations were applied to improve model robustness.

Dataset Split: 70% training, 15% validation, 15% testing. Stratified sampling was used to preserve class balance.

3.4 Model Selection and Hybrid Design

ResNet-50: Known for residual skip connections that prevent vanishing gradients, enabling deeper networks to learn effectively.

It is defined by mapping inputs to outputs via skip connections:

$y = F(x, \{Wi\}) + x$, where $F(x, \{Wi\})$

is the residual mapping (typically two or three convolutional layers) and x

is the input, bypassing the layers to solve vanishing gradient problems.

ConvNeXt: A modern CNN architecture that uses transformer-like components such as large kernel sizes and layer normalisation, enabling better representation learning.

In our hybrid approach, hybrid model (combination of both models) works on the image to get the results.

3.5 Training Procedure

Optimiser: Adam

Learning Rate: 0.001

Batch Size: 16

Epochs: 20

Loss Function: Cross-Entropy Loss

Only the classifier layer was trained, keeping feature extractor weights frozen. This reduced training time and prevented overfitting on limited data.

4 Models

4.1 ResNet-50

ResNet-50, short for Residual Network with 50 layers, is one of the most popular deep learning models for image classification. It's particularly useful when dealing with complex image data, such as satellite or remote sensing imagery. What makes ResNet-50 special is its use of residual learning, which introduces shortcut connections that allow information to skip certain layers in the network. This clever design helps the model train faster and more effectively, even when it has many layers, by preventing problems like vanishing gradients. Because of its strong architecture and reliable performance, ResNet-50 is now also popular in Land Use and Land cover (LULC) classification, where distinguishing even the smallest visual differences between types of land can have a huge effect (Fig. 1).

Why ResNet-50 for LULC?

ResNet-50 is specifically adapted to the task of Land Use and Land Cover (LULC) classification due to the capacity to represent fine-grained characteristics and complicated spatial patterns inside photographs. Numerous land cover types like urban, barren land, and agricultural fields can be visually similar yet deep architecture of ResNet-50 enables it to learn finer details that can be used to differentiate the classes more precisely. Despite its depth, the model remains efficient thanks to its residual connections, which

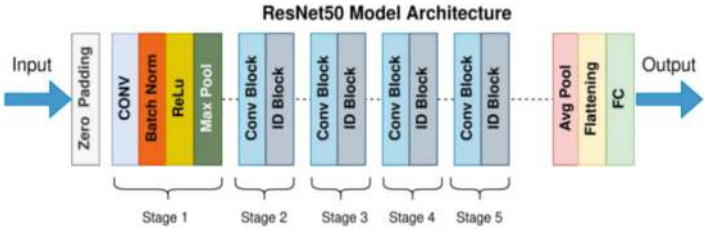


Fig. 1. This figure shows the model architecture of Resnet-50 model.

prevent issues like vanishing gradients and overfitting, allowing it to generalize well even on unseen data.

Another advantage of ResNet-50 is its adaptability across different image resolutions, making it effective for both high-resolution aerial photographs and moderate-resolution satellite imagery. This versatility ensures that the model performs consistently across varied remote sensing datasets. Moreover, by using transfer learning with pretrained weights, ResNet-50 significantly reduces the time and computational cost of training, which is especially valuable in research or resource-constrained environments. Finally, the model's strong feature extraction capability enables it to capture both local details and global spatial context—essential for interpreting complex landscapes that include a mixture of natural and man-made features such as vegetation, water bodies, roads, and built-up areas. Together, these qualities make ResNet-50 a powerful and efficient choice for LULC classification tasks.

4.2 ConvNeXt

ConvNeXt represents a modern evolution in convolutional neural network (CNN) design, introduced by Liu et al. in 2022. It builds on the strengths of traditional CNNs while incorporating innovative ideas inspired by Vision Transformers (ViTs). Essentially, ConvNeXt retains the computational efficiency and simplicity of convolutions but integrates transformer-like strategies that help it capture broader spatial relationships and improve its generalization ability. These enhancements allow the model to better understand complex visual patterns and subtle differences within images. Because of this, ConvNeXt has proven to be highly effective for Land Use and Land Cover (LULC) classification, where accurately distinguishing between similar landscape elements—such as vegetation, water bodies, and built-up regions—requires both precision and strong contextual understanding (Fig. 2).

4.2.1 Why ConvNeXt for LULC?

ConvNeXt serves as a bridge between traditional Convolutional Neural Networks (CNNs) and modern Vision Transformers (ViTs). By integrating transformer-inspired mechanisms into a convolutional framework, it effectively captures both local textures and global spatial relationships—an essential capability for interpreting the complexity of real-world terrains. One of the key strengths of ConvNeXt lies in its ability to distinguish between visually similar classes, such as water bodies and wetlands, which

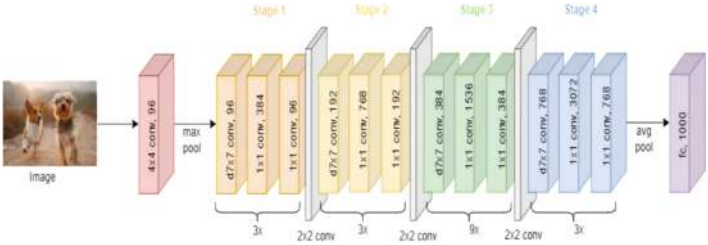


Fig. 2. This figure shows the architecture of ConvNext model.

are often difficult to separate in Land Use and Land Cover (LULC) classification. This precision reduces class confusion and enhances the overall clarity of predictions.

Moreover, ConvNeXt is designed with scalability and adaptability in mind. It performs efficiently across a range of dataset sizes, from small experimental collections to large-scale remote sensing datasets, without demanding excessive computational resources. This makes it practical for both academic research and real-world applications. Additionally, its robust architecture allows the model to generalize well across variations in image resolution, quality, and environmental conditions, maintaining reliable performance even when satellite images differ in acquisition settings. Altogether, ConvNeXt offers a balanced combination of accuracy, efficiency, and versatility, making it a strong candidate for advanced LULC classification tasks.

4.3 Hybrid Model

In order to provide more accurate Land Use and Land Cover (LULC) classification of satellite images, we have created a hybrid deep learning model which combines the features of two convolutional neural networks, which are already powerful: ResNet-50 and ConvNeXt Base. Rather than using each model independently, we combined them into a single framework that allows both networks to complement each other. This integration enables the model to learn a wider variety of visual patterns and extract richer, more detailed feature representations, ultimately leading to better classification performance across diverse land cover types.

Combining both models, we leverage ResNet-50's fine-grained feature extraction and ConvNeXt's ability to model global context, leading to more robust predictions. This ensemble-like approach reduces class-specific errors and produces smoother decision boundaries (Fig. 3).

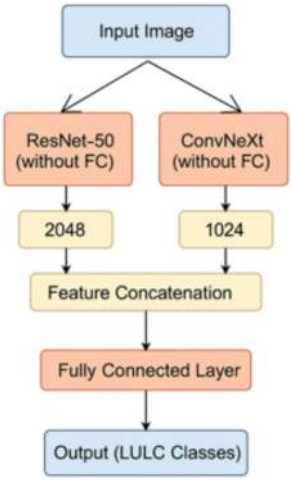


Fig. 3. This is a flowchart of proposed Hybrid Model architecture. Features from Resnet-50 and ConvNext are extracted and passed through a fully connected layer for LULC classification.

5 System Design and Implementation

5.1 System Design (Fig. 4)

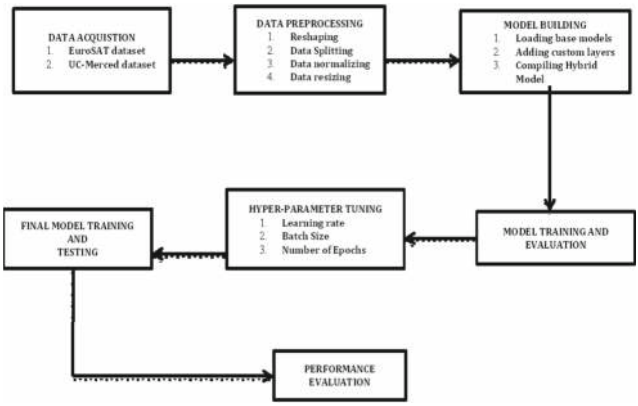


Fig. 4. Figure shows system design architecture from data collection to testing the model.

5.2 Data Acquisition

The study uses two datasets: EuroSAT and UC Merced. EuroSAT contains Sentinel-2 satellite images that cover 10 land use and land cover (LULC) classes. UC Merced includes high-resolution aerial images that cover 21 land use categories. Combining these datasets creates a diverse and challenging dataset for classification.

5.3 Dataset

EuroSAT: The proposed method collects image data from the EuroSAT dataset to classify land use and cover. This dataset is used to classify land using geospatial images. Each image is 64x64 pixels with a ground sampling distance of 10 m. All images were collected from the Sentinel-2 satellite.

UC Merced: This dataset contains land use images categorised into 21 classes, with 100 images for each class. Each image has a resolution of 256x256 pixels.

5.4 Dataset Preprocessing

Merging: To prepare the training data, images from the EuroSAT and UC Merced datasets were resized to a uniform resolution of 224x224 pixels. This ensures compatibility with deep learning models. The combined dataset was split into three subsets: 70% for training, 15% for validation, and 15% for testing.

Input Size: EuroSAT, with satellite images sized at 64x64 pixels, and UC Merced, where each image is 256x256 pixels. Since deep learning models require all input images to be the same size for efficient training and prediction, we standardised both datasets by resizing or cropping all images to a fixed dimension of 224x224 pixels.

5.5 Data Transformations

Several data augmentation techniques were applied during preprocessing to improve the model's generalisation. First, we used Random Resized Crop. Next, we applied Random Horizontal and Vertical Flips. Finally, we performed normalisation.

5.6 Model Selection

The choice of Land Use and Land Cover (LULC) classification models in this study is based on the datasets' nature, their size, and the complexity of the classification task. We selected two deep learning models, ResNet-50 and ConvNeXt, for their effectiveness in image classification.

5.7 Transfer Learning Approach

In our case, we used ResNet-50 and ConvNeXt, which were initially trained to recognise everyday objects in the ImageNet dataset. Although that dataset is very different from satellite images, the early layers of these models still extract functional patterns like shapes, edges, and textures. These features are not specific to one type of image; they can also assist with LULC classification.

Transfer learning lets us reuse existing capabilities instead of learning them from scratch. This saved training time and made our model more stable, even with limited satellite data.

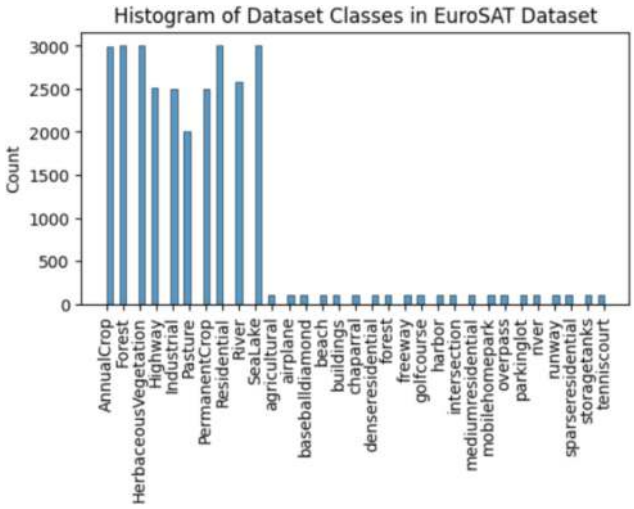


Fig. 5. This figure shows the Histogram of the model where all the classes of LULC classification are combined.

Fig. 6. This figure shows the model accuracy by hybrid model.

6 Results and Discussion

6.1 Results

The screenshots of the Land Use Land Cover classification are (Figs. 5, 6, 7, 8 and 9):

```
Evaluating Combined Dataset: 100% [██████████████████████████] 1819/1819 [07:14<00:00, 4.19it/s]

Overall Classification Report:

Class                precision  recall    f1-score  support
-----------------------------------------------------------
AnnualCrop           0.9327     0.9610    0.9466    3000
Forest               0.9604     0.9861    0.9731    3100
HerbaceousVegetation 0.9347     0.9297    0.9322    3000
Highway              0.9342     0.8796    0.9061    2600
Industrial           0.9610     0.9672    0.9641    2500
Pasture              0.9158     0.9355    0.9256    2600
PermanentCrop        0.9147     0.9132    0.9139    2500
Residential          0.9756     0.9743    0.9749    2700
River                0.9019     0.9158    0.9088    2600
SeaLake              0.9400     0.9593    0.9744    3000
agricultural         0.9901     1.0000    0.9958    100
airplane             1.0000     0.9900    0.9950    100
baseballdiamond      0.9314     0.9500    0.9406    100
beach                1.0000     1.0000    1.0000    100
buildings            0.9574     0.9000    0.9278    100
chaparral            1.0000     1.0000    1.0000    100
denseresidential     0.8641     0.8900    0.8768    100
freeway              0.9809     1.0000    0.9479    100
golfcourse           0.9804     0.9300    0.9588    100
harbor               0.9804     1.0000    0.9901    100
intersection         0.9271     0.8900    0.9082    100
mediumresidential    0.8426     0.9100    0.8758    100
mobilehomepark       1.0000     0.9700    0.9848    100
overpass             1.0000     0.8700    0.9305    100
parkinglot           0.9709     1.0000    0.9852    100
runway               0.9804     1.0000    0.9901    100
sparseresidential    0.9579     0.9100    0.9333    100
storagetanks         0.9500     0.9500    0.9500    100
tenniscourt          0.9896     0.9500    0.9694    100

accuracy                                  0.9770    29101
macro avg            0.9736     0.9794    0.9710    29101
weighted avg         0.9772     0.9770    0.9769    29101

Overall Accuracy: 0.9770
Overall Precision: 0.9772
Overall Recall: 0.9770
Overall F1-Score: 0.9769
```

Fig. 7. This figure shows the overall model classification report.

6.2 Discussion

6.2.1 Class-Wise Results

Compared to individual models, the hybrid network showed significant improvement in distinguishing between visually similar classes such as “Residential” vs. “Industrial” and “Barren” vs. “Highway.”

6.2.2 Comparative Analysis

Standalone ResNet-50: High accuracy but occasional misclassification of shrubland and agricultural land.

Standalone ConvNeXt: Better global feature extraction but slightly lower precision for smaller classes.

Hybrid Model: The Combined feature set mitigated individual weaknesses and provided more consistent classification results.

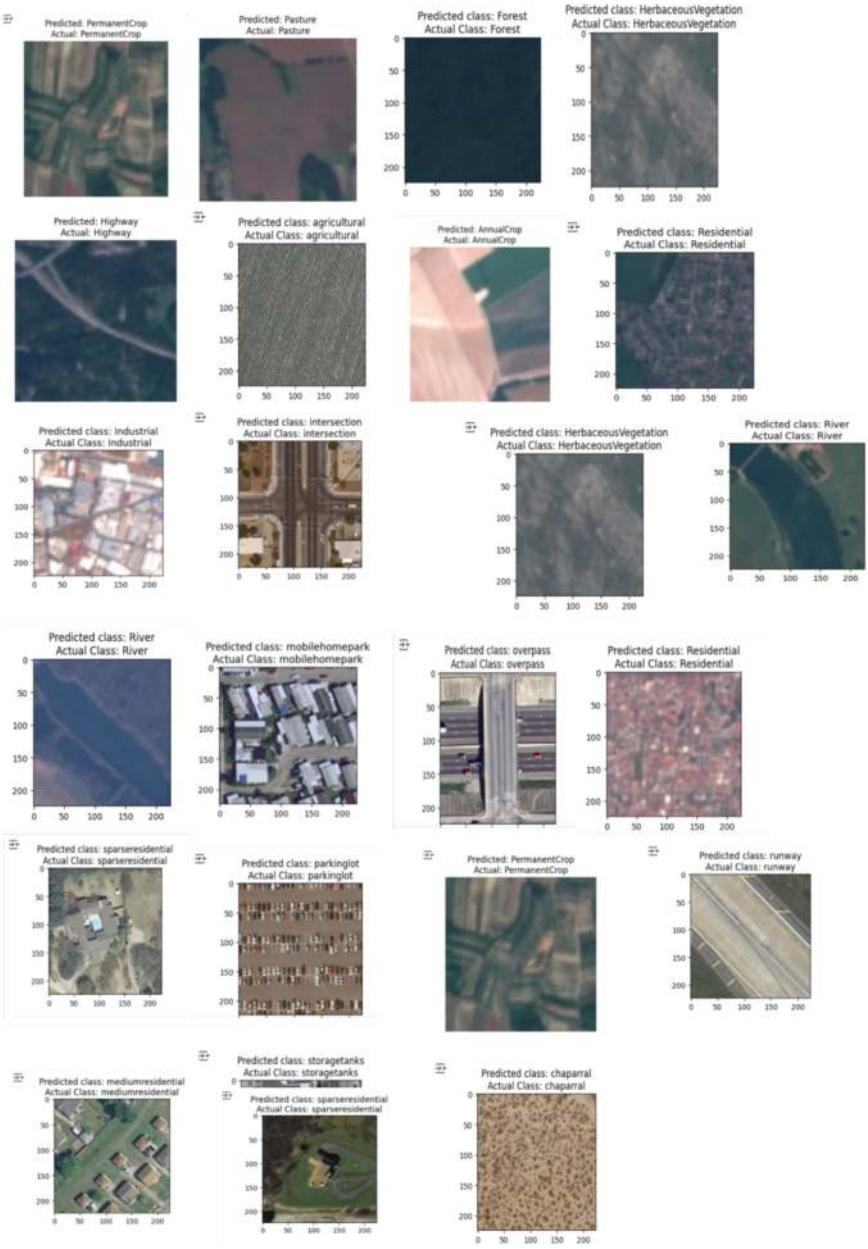


Fig. 8. This figure shows the correct prediction results by using the hybrid model.

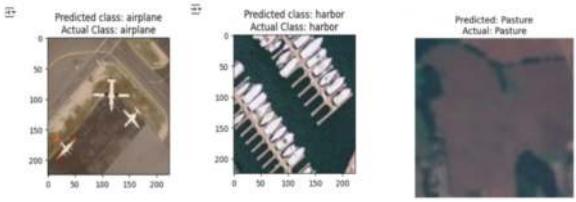


Fig. 8. (*continued*)

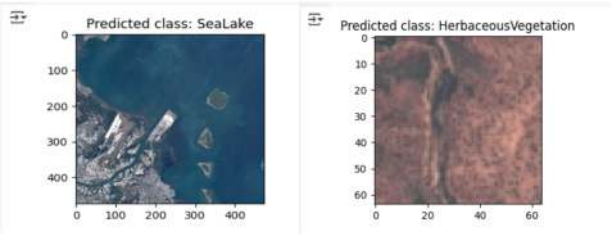


Fig. 9. This figure shows the prediction results on unseen data.

7 Limitation

Although the proposed system attained good performance, there are a number of limitations that should be established. The biggest weakness is that this study is restricted to the area of classification of static images only and it does not provide the dynamic changes in the land use and land cover which is imperative in tracking dynamic changes in the land such as urban sprawl, deforestation, and crop cycle changes. The second shortcoming is that it is computationally expensive to train deep learning models with large datasets despite transfer learning decreasing the number of trainable parameters, they still need to be trained with the help of a GPU. Also, although the merging of EuroSAT and UC Merced offers diversity to the resolution and to the representation of the classes, all the global land cover situations are not covered by it, which implies that the generalisation ability of the model will decrease under the conditions of the application to sites with different spectral signatures or land-use patterns. Others like the dataset also have a problem of imbalance in classes with certain classes having much fewer than the rest and this can affect the performance of the model on the underrepresented classes. Finally, since the study freezes the backbone layers of both ResNet-50 and ConvNeXt, the hybrid model may not fully exploit its feature representation capacity, potentially leaving performance gains untapped.

8 Future Work

Future Looking ahead, future research should focus on addressing the current limitations of this work to make the proposed system more reliable, flexible, and practical for real-world use. One promising direction is the inclusion of multi-temporal and multi-spectral data, which would allow the model to observe seasonal variations and detect

gradual environmental changes over time. This enhancement could significantly improve its usefulness in fields such as climate monitoring, agricultural planning, and disaster management.

Another potential improvement lies in fine-tuning the entire hybrid network, rather than limiting training to the classifier layer. Doing so could help the model adapt more effectively to the distinct characteristics of remote sensing imagery. In parallel, exploring transformer-based architectures—such as the Vision Transformer (ViT) or Swin Transformer—may further enhance the model's ability to capture long-range spatial relationships and complex feature dependencies.

From a practical standpoint, developing a web-based decision-support platform could make the system more accessible to policymakers, researchers, and field practitioners. Such a platform could allow users to upload satellite images, view classification outputs, generate analytical reports, and interactively explore spatial trends.

Additionally, adopting semi-supervised or self-supervised learning methods would make it possible to take advantage of the large amounts of unlabeled satellite imagery available, reducing the reliance on manual annotation. Finally, extending the framework to support real-time or near-real-time processing could transform its role in urgent applications such as disaster response, rapid land-use assessment, and environmental monitoring.

9 Conclusion

In summary, this study extends the work of the base paper by introducing a broader and more diverse dataset, along with a thoughtfully designed hybrid deep learning model. By combining two powerful architectures, the proposed model enhances its ability to learn complex visual patterns and delivers more accurate predictions across multiple evaluation metrics. The integration of both networks allows the system to balance their individual strengths—ResNet-50's deep feature extraction and ConvNeXt's strong generalization—resulting in improved performance in terms of accuracy, precision, recall, and F1-score.

A key advantage of this hybrid setup is its adaptability. Since satellite imagery often varies in resolution, source, and quality, the model's flexible design enables it to handle such variations effectively. The results demonstrate that hybrid learning can be a robust strategy for managing diverse remote sensing data while maintaining consistent performance. Furthermore, through transfer learning and fine-tuning, the model achieves strong results even with limited labeled data, showing both efficiency and scalability. Notably, these improvements were obtained in fewer training epochs, highlighting the method's computational efficiency.

Overall, the findings suggest that integrating multiple architectures in a hybrid framework offers a promising and practical approach for real-world LULC classification, combining high accuracy with efficient learning and adaptability to diverse datasets.

References

Balarabe, A.T., Jordanov, I.: LULC image classification with a convolutional neural network. In: IEEE International Geoscience and Remote Sensing Symposium (2021)

Boonpook, W., Tan, Y., et al.: Deep learning semantic segmentation for land use land cover types using Landsat 8 imagery. ISPRS Int. J. Geo Inf. **12**(1), 14 (2023)

Cahapin, E., Fajardo, A.: Fine-Tuning Res Net 50 for Land Use and Land Cover Classification in the Philippines. In: 2024 International conference on emerging smart computing and informatics (ESCI), vol. 2024, pp. 1–5, Pune, India

Harini, M., Selvavarshini, S., Narmatha, P., Anitha, V., Selvi, S.K., Manimaran, V.: Resnet-50 integrated with attention mechanism for remote sensing classification. In: Nanda, U., Tripathy, A.K., Sahoo, J.P., Sarkar, M., Li, K.C. (eds.) Lecture Notes in Networks and Systems Advances in Distributed Computing and Machine Learning. ICADCML 2024, vol. 955. Springer, Singapore (2024)

He, K., Zhang, X., Ren, S., Sun, J., (2015). Deep residual learning for image recognition ar Xiv: 1512.03385, (2015)

https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f

https://medium.com/augmented-startups/convnext-the-return-of-convolution-networks-e70cbe8dabcc

https://www.researchgate.net/figure/Land-use-land-cover-type-in-India-derived-from-MODIS-Terra-LULC-yearly-composite-data_fig1_285238507

Hu, K., Zhang, E., Dai, X., Xia, M., Zhou, F., Weng, L., Images. Remote Sens. (2022)

Irfan, G., Sun, Y.L., Zhang, H.: Cascaded deep learning model for accurate land use and land cover classification. In: IGARSS 2024–2024 IEEE International Geoscience and Remote Sensing Symposium, pp. 3077–3080, Athens, Greece (2024)

Khan, B.A., Jung, J.-W.: Semantic segmentation of aerial imagery using U-net with self-attention and separable convolutions. Appl. Sci. **14**(9), 3712 (2024)

Krivoguz, D., Chernyi, S.G., Zinchenko, E., Silkin, A., Zinchenko, A.: Using Landsat-5 for accurate historical LULC classification: a comparison of machine learning models. Data. **8**, 138 (2023)

Lin, H.: MCSGNet: an encoder–decoder architecture network for land cover classification. Remote Sens. (2023)

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., (2022) "A conv net for the 2020s," *ar Xiv*: 2201.03545 [cs], (2022)

Papoutsis, I., Bountos, N.I., Zavras, A., Michail, D., Tryfonopoulos, C.: Benchmarking and scaling of deep learning models for land cover image classification. ISPRS J. Photogramm. Remote Sens. **195** (2023)

Rengma, N.S., Yadav, M.: Generation and classification of land use and land cover datasets in the Indian states: a comparative study of machine learning and deep learning models. In: Research Square (2023)

Sertel, E.; (2022). Land Use and Land Cover Mapping Using Deep Learning Based Segmentation Approaches and VHR Worldview-3. . 305

Singh, N.J., Nongmeikapam, K.: Semantic segmentation of satellite images using deep-Unet. Arab. J. Sci. Eng. (2023)

Solórzano, J.V., Mas, J.F., Gao, Y., Gallardo-Cruz, J.A.: Land use land cover classification with U-net: advantages of combining Sentinel-1 and Sentinel-2 imagery. Remote Sens. **13**(18), 3600 (2021)

Yadav, D., et al.: Satellite image classification using a deep learning approach. Earth Sci. Inf. **17**(3), 2495–2508 (2024)

Zhao, Z., Hatamleh, W.A., et al.: Comparison of three machine learning algorithms using Google earth engine for land use land cover classification. Rangeland ecology& management. **92**, 129–137 (2024)

A Hybrid Deep Learning and Machine Learning Model for Enhancing the Accuracy of Real-Time Traffic Dynamics Forecasting

Minh Hieu Nguyen, Dmitry Ilin, and Irina Ivanova(✉)

MIREA—Russian Technological University, Moscow 119454, Russia
ivanova_i@mirea.ru

Abstract. Accurate real-time road traffic forecasting is essential for traffic management, yet many existing models struggle with the nonlinear and dynamic nature of traffic flow. The paper presents a hybrid model, Hybrid-TrafficNet, which integrates a Bi-LSTM neural network with an attention mechanism for analyzing temporal patterns in traffic flow, combined with the XGBoost method for prediction refinement. The model was tested and compared against conventional machine learning and deep learning algorithms, including Linear Regression, Decision Trees, XGBoost, SVM, ANN, and LSTM. A dataset of hourly road traffic of the I-94 highway over the period from 2012 to 2018. Forecasting accuracy was evaluated using metrics such as RMSE, MAPE, R^2, and MAE, alongside an analysis of the loss functions in deep learning models. The obtained results confirmed that the proposed hybrid approach improves the accuracy of traffic dynamics prediction compared to alternative methods. The proposed hybrid model may prove useful for intelligent transportation systems, where accurate traffic forecasting is required.

Keywords: Hybrid model · Traffic forecasting · Bi-LSTM · Attention mechanism · XGBoost · Machine learning · Deep learning

1 Introduction

Traffic dynamics forecasting is a critical task for the effective optimization of transportation systems and real-time traffic management. Recent advances in big data technologies and artificial intelligence have led to the widespread adoption of machine learning and deep learning techniques in this field [1–10]. Particular attention has been given to hybrid models that combine the strengths of different approaches, thereby enhancing the accuracy and reliability of predictions based on complex and multidimensional data [3, 7, 9, 11, 12].

This study introduces a hybrid model named Hybrid-TrafficNet, which integrates a Bi-LSTM neural network for capturing temporal dependencies in traffic flow, an attention mechanism to highlight key influencing factors [13], and the XGBoost method to refine the final predictions [14]. While LSTM-based and XGBoost-based models have been widely employed across various forecasting tasks [8], their combined application to traffic prediction problems remains relatively underexplored [4, 15, 16].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 411–418, 2027.
https://doi.org/10.1007/978-3-032-27834-0_38

The Hybrid-TrafficNet model is evaluated against conventional machine learning methods, including Linear Regression, Random Forest, Decision Trees, XGBoost, and SVM, as well as deep learning models such as ANN and LSTM. The findings demonstrate that the combined approach significantly improves prediction accuracy, particularly when dealing with diverse input factors such as weather conditions, holidays, and temporal patterns [17, 18].

To objectively assess model performance, evaluation metrics such as RMSE, MAPE, R^2, and MAE were employed, providing a comprehensive analysis of forecasting effectiveness under the influence of various external factors [8, 19, 20].

Section 2 presents the mathematical formulation of the problem and the architecture of the Hybrid-TrafficNet model, including descriptions of its core components: Bi-LSTM, the attention mechanism, and XGBoost. Section 3 provides a comparative analysis between the hybrid model and conventional machine learning and deep learning methods, followed by a discussion of the results. Finally, Sect. 4 provides conclusions and directions for future work.

2 Hybrid-TrafficNet: a Hybrid Model for Traffic Forecasting

To improve the accuracy of traffic flow forecasting, a hybrid model Hybrid-TrafficNet is proposed, combining deep learning and machine learning algorithms. The core components of the model include a Bi-LSTM network for capturing temporal dependencies, the XGBoost algorithm for modeling nonlinear relationships between features, and an attention mechanism that enables the model to focus on key factors influencing traffic dynamics [2, 4, 11].

The task of forecasting traffic flow intensity is approached as a typical time series prediction problem. The objective is to determine the traffic flow value Y_t at a given time t, based on a set of input variables and contextual factors that incorporate both historical information and current conditions [1]. The input set X_t includes parameters characterizing traffic patterns, meteorological indicators, and holiday data within the time window from $t - n$ to the current time t. Additionally, a set of factors S_t is considered, representing the influence of sudden events, weather changes, or mass events that may significantly affect traffic dynamics [15]. Eq. (1) presents the mathematical formulation of the problem.

$$Y_t = f\left(X_{t-n}, X_{t-n+1}, \ldots, X_t, S_t\right) \tag{1}$$

where the function $f(\cdot)$ is modeled using the hybrid architecture Hybrid-TrafficNet, which integrates heterogeneous data to ensure high forecasting accuracy.

As shown in Fig. 1, the Hybrid-TrafficNet model comprises three core components.

First, the Bi-LSTM module [15] captures temporal dependencies within the data by processing sequences in both forward and backward directions, enabling the model to account for complete dynamic relationships. It processes sequential data and generates the output state h_{i}.

Second, the attention mechanism [11] highlights the most influential features (such as time, weather, and holidays), thereby enhancing both the interpretability and predictive

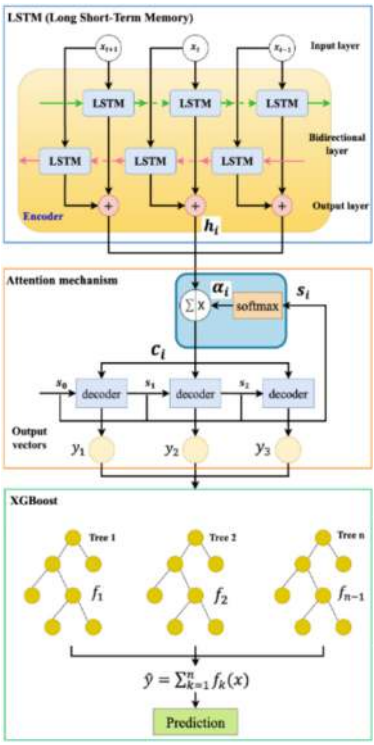


Fig. 1. Architecture of the Hybrid-TrafficNet Model.

accuracy of the model. The attention weights a_i are used to compute the context vector c_i, which aggregates information from all hidden states of the Bi-LSTM at each time step. After the context vector c_i is computed, it is integrated with the hidden state h_i from the Bi-LSTM to enhance the decoding process. This integration takes place in the decoder, which transforms the combined input into the final prediction y_i for each time step.

Third, the XGBoost [4] component models nonlinear dependencies through boosted decision trees, offering robustness against overfitting and achieving high prediction accuracy. Each tree, denoted as $f_k(x)$, makes a prediction based on the input features x, and the final predicted value is obtained by summing the contributions of all trees in the model.

Bi-LSTM extends the classical LSTM architecture by processing dŷata sequences in both directions forward (past to future) and backward (future to past) [15]. This bidirectional structure enables the model to capture dependencies between time steps from both sides, thereby improving the accuracy of time series analysis and enhancing the recognition of complex patterns within the data.

3 Evaluation of the Hybrid-TrafficNet Model

For model training and evaluation, a traffic dataset was used, collected from road sensors installed on the I-94 highway between Minneapolis and Saint Paul over the period from 2012 to 2018. The data include information on traffic density, temperature, weather conditions, time of day, day of the week, and public holidays. These variables enable the model to account for a wide range of factors influencing traffic dynamics, thereby supporting more accurate real-time forecasts.

After training the model, it is essential to evaluate its performance to determine how accurately it predicts traffic intensity. The evaluation is conducted using regression metrics [19, 20], as the target variable traffic intensity is continuous in nature.

Model evaluation is conducted using various regression metrics such as RMSE, MAPE, MAE, and R^2, which enable a detailed assessment of prediction accuracy and the model's ability to explain variations in the target variable. Each of these metrics plays a crucial role in understanding the overall quality of the model and its applicability to real-world traffic dynamics forecasting tasks.

The training set includes data from the period 2012 to 2017, accounting for approximately 82% of the total dataset, while the test set comprises data from 2018, representing about 18%. This temporal split enables an objective evaluation of the models' forecasting capabilities on independent data, taking into account seasonal and cyclical patterns within the time series.

In the experimental framework, the hybrid Hybrid-TrafficNet model was compared with traditional machine learning methods (Linear Regression, Random Forest, Decision Trees, XGBoost, SVM), as well as deep learning models (ANN, LSTM).

The forecasting task was implemented using the direct multi-step prediction approach, experimenting with a forecasting horizon of six steps ahead, enabling the simultaneous prediction of multiple future values. In addition, a 24-step horizon was also applied to capture full-day traffic dynamics. The model receives structured contextual data as input for traffic volume prediction. Unlike recursive strategies, this method directly computes each future value from the input data, reducing error accumulation and improving prediction accuracy over longer horizons [21]. This approach aligns with modern standards for multi-step time series forecasting and has proven effective in deep neural network architectures for traffic prediction tasks.

Figure 2 presents comparative loss curves during training and validation for three models: Hybrid-TrafficNet, LSTM, and ANN. The Hybrid-TrafficNet model demonstrates a rapid decrease in loss from the beginning of training and continues to converge steadily, indicating efficient learning. The LSTM model reduces loss more slowly but maintains improvement throughout training, although its final loss value remains slightly higher than that of Hybrid-TrafficNet. The ANN model shows a very slow decrease in loss, with little to no improvement after around 20 epochs, suggesting insufficient optimization compared to the other two models. Overall, Hybrid-TrafficNet achieves the best performance in terms of loss reduction, followed by LSTM, while ANN demonstrates the least effectiveness. As shown, the Hybrid-TrafficNet model exhibits lower validation losses compared to LSTM and ANN, particularly during the initial epochs. While LSTM and ANN display a gradual decrease in validation loss, Hybrid-TrafficNet

achieves superior performance in fewer epochs, indicating faster convergence and a stronger generalization capability.

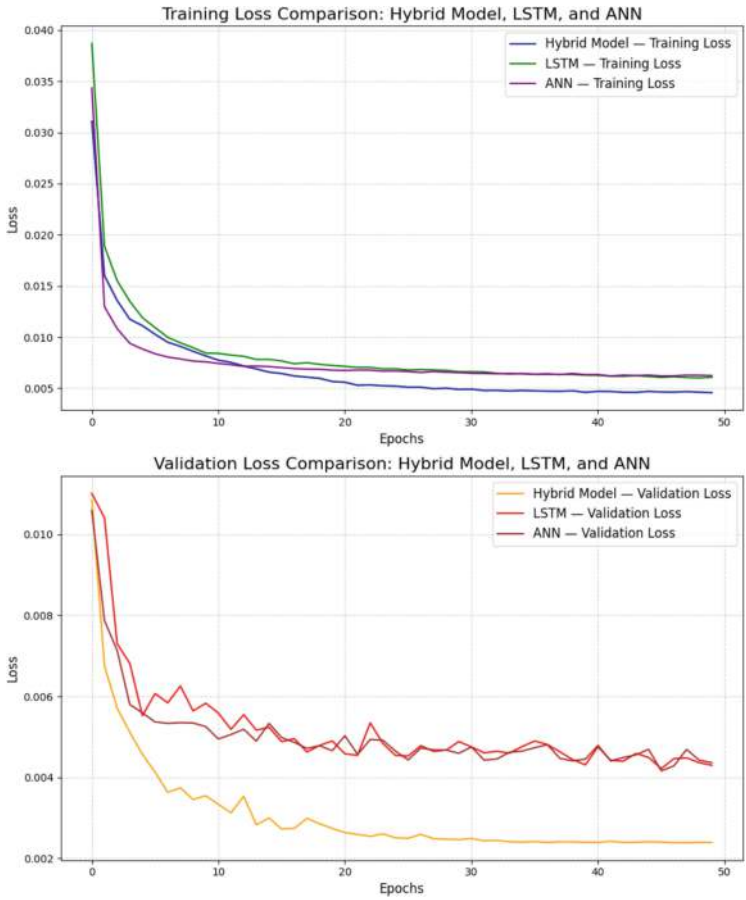


Fig. 2. Comparison of training and validation loss across different models.

The evaluation metrics were calculated on the normalized test dataset to ensure fairness in evaluation. The results, summarized in Table 1, reflect the overall predictive accuracy of each model throughout the entire forecasting period, thereby facilitating the selection of the most suitable approach for the traffic volume prediction task [16, 18]. Model evaluation results are rounded to four decimal places. The metrics are calculated on normalized data (range [0, 1]) using MinMaxScaler to ensure a fair comparison between models.

Figure 3 provides a comparative analysis between the predicted traffic volumes generated by the Hybrid-TrafficNet model and the actual observations from the 2018 test dataset. Each subplot illustrates a six-hour forecasting horizon, following the direct multi-step prediction strategy, in which six consecutive future values are simultaneously

Table 1. Comparison of models.

		RMSE	MAPE	R^2	MAE
Machine learning	Linear regression	0,1653	18,4234	0,6763	0,1438
	Random forest	0,0578	6,4655	0,9157	0,0502
	Decision trees	0,0746	8,6976	0,8723	0,0623
	XGBoost	0,0641	7,0512	0,8993	0,0522
	SVM	0,1208	13,5863	0,7341	0,0914
Deep learning	LSTM	0,0709	7,8936	0,8867	0,0611
	ANN	0,0787	10,2276	0,8691	0,0673
Hybrid model	Hybrid-Trafficnet	0,0424	5,5258	0,9423	0,0320

estimated. The model demonstrates a high level of agreement with the observed traffic data, capturing both gradual variations and sudden peaks in volume. Although minor discrepancies occur in some intervals, they are largely attributed to external factors not represented in the data, such as traffic incidents or infrastructural disruptions. Collectively, these results confirm the robustness and short-term predictive accuracy of the proposed approach in realistic traffic forecasting scenarios.

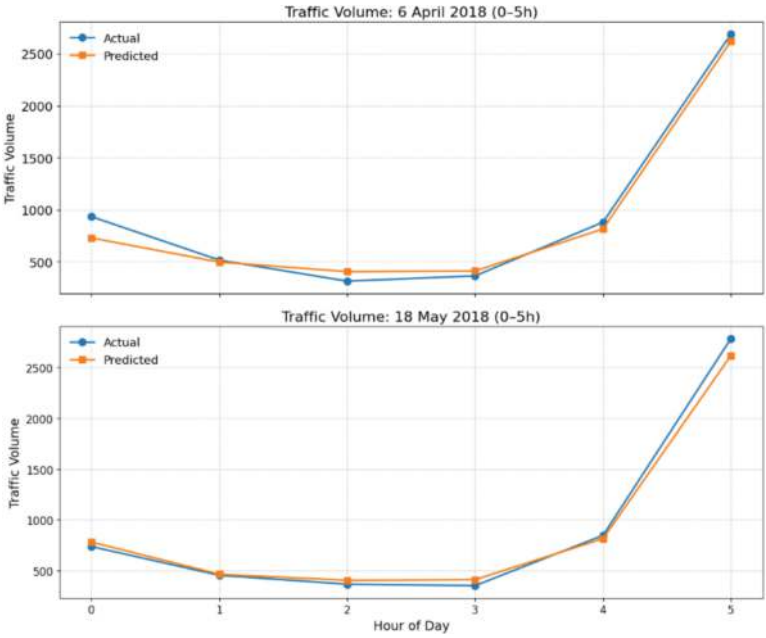


Fig. 3. Six-hour traffic forecast performance.

The comparative analysis of different models showed that Hybrid-TrafficNet achieved the best results across all key metrics, including RMSE, MAPE, R^2, and MAE. By combining Bi-LSTM-based feature extraction with an attention mechanism and XGBoost regression and employing a direct multi-step prediction strategy, the model accurately forecasted multiple future values simultaneously. While Random Forest, LSTM, and XGBoost also performed well, Hybrid-TrafficNet demonstrated superior accuracy and generalization. In contrast, Linear Regression and SVM lagged behind, confirming their limited effectiveness in modeling complex traffic patterns.

4 Conclusion

The Hybrid-TrafficNet model exhibited strong overall forecasting performance, achieving better results across all evaluation metrics with RMSE—0.0424, MAPE—5.5258%, R^2—0.9423, and MAE—0.0320 compared to the number of machine learning and deep learning models. These results show the effectiveness of combining machine learning and deep learning with attention mechanisms, enabling more precise predictions than traditional or standalone models. The findings indicate that the combination of attention mechanisms and ensemble methods may enhance forecasting performance.

Future research can be aimed at extending the forecasting horizon. The integration of real-time external data such as live traffic incidents and infrastructure changes can also be explored to improve contextual awareness. Furthermore, research efforts can be focused on optimizing the model for deployment in edge computing environments to support low-latency applications. Finally, adaptive learning techniques can be investigated to allow the system to evolve dynamically with incoming traffic data, enhancing its robustness and long-term applicability.

References

1. Gomes, B., Coelho, J., Aidos, H.: A survey on traffic flow prediction and classification. Intell. Syst. Appl. **20**, 200268 (2023)
2. Wang, J. et al.: SA-bi-LSTM: self-attention with bi-directional LSTM-based intelligent model for accurate fake news detection to ensure information integrity on social media platforms. IEEE Access. **12**, 48436–48452 (2024)
3. Lou, J., Cui, L., Li, Y.: Bi-LSTM price prediction based on attention mechanism. https://arxiv.org/abs/2212.03443 (2022)
4. Zheng, H., Wu, Y.: A XGBoost model with weather similarity analysis and feature engineering for short-term wind power forecasting. Appl. Sci. **9**(15), 3019 (2019)
5. Kowsher, M. et al.: LSTM-ANN & BiLSTM-ANN: hybrid deep learning models for enhanced classification accuracy. Procedia Comput. Sci. **193**, 131–140 (2021)
6. Węglarczyk, S.: Kernel density estimation and its application. ITM Web Conf. **23**, 00037 (2018)
7. Deekshetha, H. R., Madhav, A. V. S., Tyagi, A. K.: Traffic prediction using machine learning. In: Suma, V., Fernando, X., Du, KL., Wang, H. (eds.) Evolutionary Computing and Mobile Sustainable Networks. Lecture Notes on Data Engineering and Communications Technologies, vol. 116, pp. 969–983. Springer, Singapore. (2022).
8. Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.-Y.: Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. **16**(2), 865–873 (2015)

9. Zhang, J., Zheng, Y., Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. Proceedings of the AAAI Conference on Artificial Intelligence. **31**(1), 1655–1661 (2017)
10. Jiang, L., Zhang, T., Zuo, Q., Tian, C., Chan, G. P., Chan, W. K. V.: A deep learning framework for traffic data imputation considering spatiotemporal dependencies. In: Proceedings of the 2022 IEEE 7th International Conference on Intelligent Transportation Engineering (ICITE), pp. 14–19. IEEE, Beijing (2022)
11. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 606–615. Association for Computational Linguistics, Austin (2016)
12. Shaheen, H., Agarwal, S., Ranjan, P.: MinMaxScaler binary PSO for feature selection. In: Proceedings of the First International Conference on Sustainable Technologies for Computational Intelligence, pp. 705–716. Springer, Singapore (2020)
13. Vaswani, A. et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., Long Beach (2017)
14. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 785–794. Association for Computing Machinery, San Francisco (2016)
15. Cui, Z., Ke, R., Pu, Z., Wang, Y.: Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. https://arxiv.org/abs/1801.02143 (2018)
16. Hu, Y.: Comparison and analysis of the effectiveness of linear regression, decision tree, and random forest models for health insurance premium forecasting. Adv. Econ., Manag. Polit. Sci. **79**, 347–353 (2024)
17. Wen, X., Li, W.: Time series prediction based on LSTM-attention-LSTM model. IEEE Access. **11**, 48322–48331 (2023)
18. Chen, Y.: Research on the prediction of Boston house price based on linear regression, random forest, XGBoost and SVM models. Highlights Bus., Econ. Manag. **21**, 27–37 (2023)
19. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. **7**, e623 (2021)
20. Hodson, T.O.: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. **15**(14), 5481–5487 (2022)
21. Pashshoev, B., Petrusevich, D.A.: Neural network analysis in time series forecasting. Russ. Technol. J. **12**(4), 106–116 (2024)

Development of a Hardware and Software System for Agroclimatic Conditions Monitoring and Machine Learning Datasets Generating

Nikolay Oleinikov[1](✉), Yuriy Grishin[2], Anatoliy Kazak[1], and Anastasia Bobarykina[1]

[1] V.I. Vernadsky Crimean Federal University, 295007 Simferopol, Russia
Oleinikov1@mail.ru

[2] Magarach All-Russia National Research Institute for Viticulture and Winemaking, Yalta, Russia

Abstract. This study analyzes the practical implementation of a real-time climate data monitoring system using IoT devices. An application programming interface (API) system was developed, combining an ESP32-based hardware platform with an ASP.NET Core server, ensuring high performance. The API enables weather monitoring, centralized data storage, and data structuring for machine learning (ML). Algorithms for validation and data export to CSV/JSON were implemented, enabling the creation of representative datasets for training neural networks.

Keywords: IoT · Climate sensors · Automated data collection · API for environmental monitoring · Microcontrollers in agrometeorology

1 Introduction

Viticulture occupies a key position in Russia's agricultural sector, exerting a growing influence on the country's economy. According to Rosstat, since 2022, the share of grapes in total agricultural output has been steadily increasing, having increased more than fourfold compared to previous years.

The primary use of grapes in Russia is the wine industry, which accounts for over 80% of the country's grape production. Sustainable viticulture development is linked to a wide range of intersectoral links with industries such as food (wine, confectionery, and canning), chemicals (use of mineral and organic fertilizers), and mechanical engineering (production of equipment for transporting and processing grapes), as well as the implementation of scientometric artificial intelligence technologies. Consumer quality indicators (tasting properties) and the antioxidant activity of wines are largely determined by the chemical composition and physiological state of the grapes, which are largely dependent on terroir conditions [1]. Among the chemical factors of grapes, phenolic compounds have the greatest impact on the oenological characteristics of wines. They determine the color and color variety of wines, as well as the aromatic profile, flavor

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 419–428, 2027.
https://doi.org/10.1007/978-3-032-27834-0_39

characteristics, and antioxidant activity of wines [2]. The biosynthesis and accumulation of phenolic compounds are inextricably linked to environmental conditions [3].

The main areas for the implementation of artificial intelligence technologies are monitoring the growing conditions of grape plants, as well as the formation of decisions based on dataset analysis regarding the ripening of grapes, the timing of harvesting, and subsequent processing methods.

Climate conditions have a decisive influence on the growth and development of plants. Air temperature and humidity, soil moisture, light levels, atmospheric pressure, and other meteorological parameters directly influence photosynthesis, water balance, and growing season.

Climate factors such as temperature, precipitation, and solar insolation duration are critical determinants affecting the chemical composition of grapes. Elevated temperatures accelerate grape ripening and potentially accelerate the onset of phenolic maturity. Deviations from optimal values for these determinants can cause yield reductions [4]. Phenolic compounds perform a protective function in plants, shielding them from the negative effects of ultraviolet radiation, especially in regions exposed to high levels of solar activity, and from pathogens. Studies have shown a significant correlation between the content of phenolic compounds and the antioxidant activity of wines [5–8]. Phenolic compounds found in grapes and grape products exhibit a wide range of biologically valuable properties (anti-inflammatory, anti-cancer) and cardiovascular protective effects [9].

Effective forecasting and decision-making regarding plant maturity, as well as managing agronomic processes, require an accurate assessment of climatic conditions in the growing area. Existing weather stations are often located at significant distances from agricultural fields, and therefore their data may be inaccurate. This limits the applicability of weather station information for precision farming. Manual measurements are labor-intensive, error-prone, and do not provide continuous monitoring.

The Internet of Things (IoT) opens up new possibilities for environmental and agrometeorological monitoring. IoT devices enable the deployment of distributed sensor networks. They autonomously collect, process, and transmit data. Such systems are characterized by low cost, flexible configuration, and scalability. Microcontrollers with wireless communication support are becoming the foundation of compact weather stations. They are capable of operating in remote areas and provide sufficient measurement accuracy for scientific and applied applications.

The use of server technologies allows for centralized processing of data received from microcontrollers. An application programming interface (API) ensures standardized data receipt and transmission.

Neural networks are an effective tool for predicting agroclimatic conditions. They are capable of identifying nonlinear relationships in large volumes of data. Training such models requires a high-quality and representative dataset. The data must be regular, reliable, and time- and spatially specific. Existing open meteorological databases often lack information at the level of a specific field or greenhouse.

The development of a specialized data collection system addresses this issue. The proposed solution combines hardware and software components.

2 Related Work Analysis

An analysis of modern scientific publications on climate monitoring in agriculture demonstrates a steady transition from centralized weather stations to distributed networks of IoT devices. Modern solutions are increasingly built on low-cost microcontrollers such as the ESP32, STM32, and Arduino microcontrollers. These platforms provide sufficient computing power, support for wireless interfaces, and low power consumption [10–14].

The most common solution is the ESP32 platform. It combines low power consumption and high performance. Built-in Wi-Fi and Bluetooth modules enable wireless communication. The microcontroller supports deep sleep modes, which extends battery life. The STM32 is used in industrial IoT devices. The series is characterized by high reliability and flexible configuration. However, it requires a more complex tool base. Arduino is also used in agricultural IoT systems. The platform has a simple architecture and an intuitive development environment. Arduino is suitable for initial prototypes and educational projects. However, it is limited in memory and processing power. The lack of built-in Wi-Fi requires additional modules.

Data transmission is carried out via various protocols [15, 16]. Wi-Fi is used in conditions with stable internet coverage, for example, in greenhouses or near fields. LoRa, LoRaWAN, and GSM are used in remote areas due to their long range and low power consumption.

The server component of most systems is implemented using modern web technologies. Popular solutions based on Python (Flask, Django), Node.js, and. NET. PostgreSQL, MySQL, Firebase, and InfluxDB are used as database management systems (DBMS), providing reliable storage of time series. Data visualization is implemented through web interfaces built using Grafana, React, or HTML forms. This allows users to monitor climate parameters in real time.

A key element of modern systems is the API. A number of the studies reviewed implement RESTful interfaces for data access. This enables integration with mobile applications, analytical platforms, and external services. However, the vast majority of APIs are focused on visualization and monitoring rather than preparing data for machine learning. There is a lack of standardization of formats, timestamps, response structures, and preprocessing methods. Some researchers consider dataset formation a separate task. A number of studies have already integrated server-side data preprocessing elements, implementing outlier filtering, missing value compensation, and aggregation across time windows.

Machine Learning (ML) is actively used to process agricultural data [17–19]. Algorithms are trained using historical and current measurements. Neural networks are particularly effective for time series analysis. They enable forecasting weather conditions, soil moisture levels, and disease risks. Classification models can recognize plant health from images. They detect signs of stress, nutrient deficiencies, and pests.

ML-based systems can also be used to optimize irrigation and fertilization. The systems make decisions based on available data. This reduces resource consumption and increases crop yields. Fuzzy Logic is used as an alternative or complement to ML [20–22]. It works well when data is insufficient.

With the development of IoT in agriculture, the need for unified solutions is increasing. Frameworks enable standardization of system development, deployment, and management. They ensure the integration of hardware and software components [23, 24]. Modern frameworks integrate sensors, microcontrollers, communication protocols, and server platforms. Frameworks also ensure modularity and scalability. Cloud platforms can provide ready-made frameworks. These solutions ensure secure data transfer and storage, as well as APIs. However, they require a constant internet connection and can be expensive.

The proposed system combines a reliable ESP32-based hardware platform with an ASP.NET Core server, ensuring high performance and security. The API is designed not only for monitoring but also for structured access to data in a format suitable for machine learning (ML). Validation, normalization, and CSV/JSON export mechanisms are implemented. This allows for the creation of representative datasets for training neural networks, for example, in tasks such as microclimate forecasting, plant stress assessment, or irrigation optimization.

3 Materials and Methods

3.1 Server-Side Application

The cross-platform ASP.NET framework was used for development. The platform supports building RESTful services with minimal overhead. Built-in support for asynchronous operations allows for efficient processing of multiple concurrent requests from IoT devices. Integration with the Entity Framework simplifies working with databases. A built-in routing and data validation system improves the reliability of the API.

MySQL was used as the database. Entity Framework was used as the ORM. It allows the developer to work with C# objects instead of manually writing SQL queries, reducing the likelihood of errors. It also supports asynchronous operations. Visual Studio Community 2022 was used as the IDE.

3.2 Microcontroller-Based Hardware and Software Module

The ESP-32-Wrover-Dev was chosen as the microcontroller board. The advantages of the ESP-32-Wrover-Dev include:

- Built-in Wi-Fi and Bluetooth, ensuring reliable wireless communication;
- A built-in camera;
- High computing power and PI Flash: 32 Mbit/RAM: Internal 520 KB + External 4 MPSRAM, allowing for data storage or camera image transmission;
- Support for power-saving modes, extending battery life;
- Simplified code development thanks to support for the Arduino IDE and ESP-IDF.

The Arduino IDE was used as the IDE.

Figure 1 shows model of the system's main components. The system consists of three groups of components:

- a microcontroller and its connected sensors;

– an API containing enpoint, an ORM database query handler;
– a MySQL database.

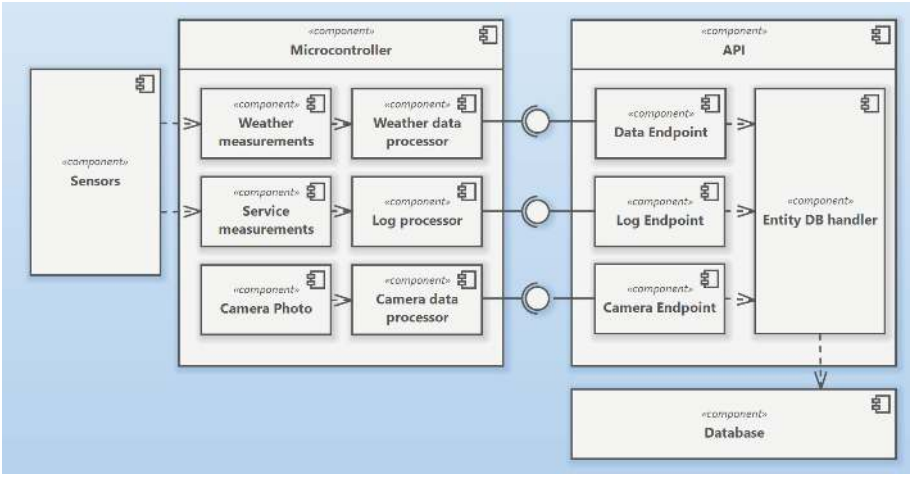


Fig. 1. Enlarged structure of the system.

4 Server-Side Application

4.1 Database Structure

Figure 2 shows a fragment of a database for storing weather data.

The Controllers table stores information about registered microcontrollers, including their model, network address (IP), geographic location (longitude and latitude), and a link to the region in which they are located.

The Regions table contains data about the regions (geographical or logical zones) in which the microcontrollers are located. This allows controllers to be grouped for easier management and data analysis.

The SensorData table stores sensor readings transmitted by the microcontrollers. It includes information about temperature, air and soil humidity, atmospheric pressure, illumination, as well as the measurement timestamp and a link to a specific microcontroller.

The Logs table stores logs sent by the microcontrollers. Each entry contains the message text, timestamp, and ID of the microcontroller that sent it.

LogCodes stores standardized message/error types that microcontrollers can send. For example, “Sensor Error” or “Low Battery.”

Camera Images stores information about images received from microcontroller cameras. The images themselves are stored separately in the file system, and the database stores the file name, its path in the file system, its size, and the timestamp of receipt, as well as a link to the microcontroller that sent the image. This implementation allows for a smaller database size and improved performance.

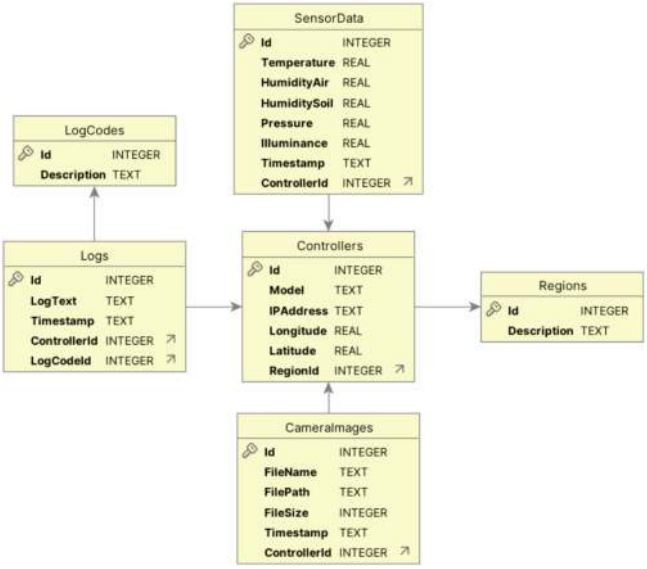


Fig. 2. A database fragment for storing data from microcontrollers.

5 API Structure

Figure 3 shows structure diagram of the API.

Input and output data are implemented using DTOs (Data Transfer Objects) using the Separation of Concerns pattern. The primary purpose of DTOs is to act as a strongly typed “intermediary” between microcontrollers, API clients, and the application’s internal logic. This implementation allows the application’s data model to be isolated from the data format used in the API, ensuring flexibility, security, and ease of support.

The DTOs are clearly divided into two groups: InputDTOs and OutputDTOs.

InputDTOs contain classes designed exclusively for receiving data from API clients, in this case, from microcontrollers. These classes define the contract that the client must adhere to when sending POST requests.

OutputDTOs contain classes designed exclusively for sending data to API clients. These classes define the structure of responses that the server returns in response to GET requests. EntityData represents the core of the data access layer in the project

architecture. Its main component is ApplicationDbContext, a central class that manages interaction with the MySQL database via Entity Framework Core.

Entities contain entity classes that directly represent tables in the database. Each entity models a specific domain object: a microcontroller, its readings, location, logs, and images. A key feature of these classes are navigation properties that implement one-to-many and many-to-one relationships. This data model ensures data integrity and enables complex database queries, reflecting real-world relationships between system objects.

Handlers, Validators, and Services comprise the API's business logic layer.

Validators are responsible for ensuring the integrity of incoming data. They use the Fluent Validation library to implement strict validation of all incoming data. Each Input-DTO has a corresponding validator that automatically checks requests for compliance with rules: value ranges (temperature, humidity), formats (IP address), required fields, etc. This check occurs early in request processing, allowing for the immediate rejection of invalid data with a 400 Bad Request error code, thereby protecting the application's underlying layers and the database from corruption.

Services is the long-term business logic and infrastructure layer. It contains classes that are not tied to a single specific HTTP request. For example, the IFile Storage Service and its implementation, File Storage Service, encapsulate file system logic, such as storing uploaded images, providing abstraction and simplifying testing. This layer ensures code reuse and independence from the HTTP protocol.

Handlers are the entry point for business logic for each HTTP endpoint. Handlers are asynchronous methods that are called when a request arrives. They receive validated data (InputDTO), interact with Services (for example, to save a file), create Entities for the database context (DbContext) for saving, and generate the final response to the client. Thus, Handlers coordinate the work of all previous layers to perform a specific business operation.

Endpoints and Extensions form the external API interface, acting as a façade through which clients (microcontrollers, users) interact with the application's internal logic. They implement the Minimal API pattern, typical of modern ASP.NET Core applications. Each route is assigned a corresponding Handler. This allows the logic of route registration to be separated from the logic of route processing.

6 Results and Conclusions

The research resulted in a high-performance API system based on an ESP32 hardware platform and an ASP.NET Core server. Real-time data allows for the rapid monitoring of climatic conditions for grapes or any other agricultural crop, allowing for timely decisions on agricultural practices to ensure a consistent harvest quality and quantity.

A server API was developed for collecting and storing data from a distributed network of microcontrollers. The project included the development and implementation of a modular software architecture based on the principles of separation of concerns, including layers of data transfer objects (DTOs), database entities, business logic (services and handlers), and validation.

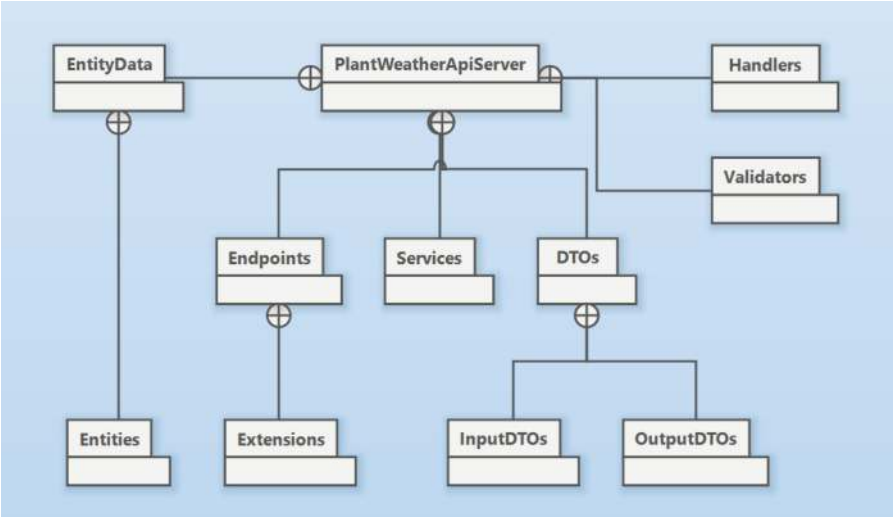


Fig. 3. API structure.

The developed API can help with range of tasks related to the centralized collection, structuring, and storage of climatic data, including temperature, humidity, atmospheric pressure, illumination, geographic coordinates, images, and system logs from devices. This provides a reliable, scalable, and secure platform for monitoring plant and environmental conditions in real time. The use of neural networks trained on real data will ensure the production of products with the required values of controlled indicators and antioxidant activity.

Acknowledgments. The study is supported by the Russian Science Foundation, Grant No. 24-25-20001.

References

1. Bai, S. et al.: Flavonoids profile and antioxidant capacity of four wine grape cultivars and their wines grown in the Turpan Basin of China, the hottest wine region in the world. Food Chem.: X. **26**, 102301 (2025). https://doi.org/10.1016/j.fochx.2025.102301
2. Teixeira, A., Eiras-Dias, J., Castellarin, S.D., Geros, H.: Berry phenolics of grapevine under challenging environments. Int. J. Mol. Sci. **14**(9), 18711–18739 (2013). https://doi.org/10.3390/ijms140918711
3. Yue, X. et al.: Effects of leaf removal on the accumulation of anthocyanins and the expression of anthocyanin biosynthetic genes in cabernet sauvignon (Vitis vinifera L.) grapes. J. Sci. Food Agric. **101**(8), 3214–3224 (2021). https://doi.org/10.1002/jsfa.10951
4. Rodriguez-Lorenzo, M. et al.: The flavour of grape colour: anthocyanin content tunes aroma precursor composition by altering the berry microenvironment. J. Exp. Bot. **74**(20), 6369–6390 (2023). https://doi.org/10.1093/jxb/erad223

5. Mavric-Scholze, E. et al.: Comparative analysis of antioxidant activity and content of (poly) phenolic compounds in cabernet sauvignon and merlot wines of Slovenian and Serbian vineyards. Food Chem.: X. **25**, 102108 (2024). https://doi.org/10.1016/j.fochx.2024.102108
6. Yang, Y., Ye, Z., Araujo, L.D., Rutan, T., Deed, R.C., Kilmartin, P.A.: Inter-regional characterisation of New Zealand pinot noir wines: correlation between wine colour, monomeric and polymeric phenolics, tannin composition, antioxidant capacity, and sensory attributes. Food Chem.: X. **467**, 142311 (2025). https://doi.org/10.1016/j.foodchem.2024.142311
7. Moravcova, P. et al.: The profiling of phenolic compounds in archive Tokaj wines using liquid chromatography and antioxidant activity analysis. Food Chem.: X. **30**, 102897 (2025). https://doi.org/10.1016/j.fochx.2025.102897
8. Ramirez-Rodrigues, M.M., Perez-Alva, A., Garcia-Torres, R.: Phenolic profile and antioxidant capacity of skin-contact white wines. Curr. Nutr. Food Sci. **21**(4), 508–517 (2025). https://doi.org/10.2174/0115734013288561240903113915
9. Kennedy, J.A., Saucier, C., Glories, Y.: Grape and wine phenolics: history and perspective. Am. J. Enol. Vitic. **57**(3), 239–248 (2006). https://doi.org/10.5344/aev.2006.57.3.239
10. AL-Nabi, N.R.A., Mayyahi, L.A.K., Majeed, A.S., Abdulnabi, A.R., Abdul-Rahaim, L.A.: Design and Implementation of a Low-cost IoT Smart Weather Station Framework. Journal of Internet Services and Information. Security. **14**, 133–144 (2024). https://doi.org/10.58346/JISIS.2024.I2.009
11. Ibn Dahou Idrissi, A., Abouabdillah, A., Chikhaoui, M. and Bouabid, R., Low-cost IoT-based monitoring system for precision agriculture E3S Web Conf. 492 01003 (2024) https://doi.org/10.1051/e3sconf/202449201003
12. Sharipah, D., Muhammad, H, Abdul Rahman, R IoT-Based Smart Agriculture Monitoring System (2025)
13. Le Van, T., Thi, P.N., Minh, P.V., Do Thi, L., Le Cong, H., Phuong, T.N.T.: Design and implementation of a wireless sensor network for smart greenhouse controller. Comm IT (Commun. Inf. Technol.) Journal. **16**, 1–8 (2022). https://doi.org/10.21512/commit.v16i1.7288
14. Sujin, J.S., Murugan, R., Nagarjun, M., Praveen, A.K.: IOT based greenhouse monitoring and controlling system. J. Phys. Conf. Ser. **1916**, 012062 (2021). https://doi.org/10.1088/1742-6596/1916/1/012062
15. Adegboye, M.: Wireless Sensor Networks Based-Internet of Thing for Agro-Climatic Parameters, p. 7.240-252 (2017)
16. Catania, P. & Vallone, Mariangela & Lo Re, Giuseppe & Ortolani, Marco. (2013). A wireless sensor network for vineyard management in Sicily (Italy). Agric. Eng. Int. CIGR J. 15.139–146.
17. Goap, A., Sharma, D., Shukla, A.K., Krishna, C.R.: An IoT based smart irrigation management system using machine learning and open source technologies. Comput. Electron. Agric. **155**, 41–49., ISSN 0168-1699, (2018). https://doi.org/10.1016/j.compag.2018.09.040
18. Tzerakis, K., Psarras, G., Kourgialas, N.N.: Developing an open-source IoT platform for optimal irrigation scheduling and decision-making: implementation at olive grove parcels. Water. **15**, 1739 (2023). https://doi.org/10.3390/w15091739
19. Wenshuang, D., Jin, P., Jin, W., Smart farm digital twin model based on edge-cloud architecture for tomato monitoring and detection, Smart Agricultural Technology: 101254. ISSN. **12**, 2772–3755 (2025). https://doi.org/10.1016/j.atech.2025.101254
20. Ramakrishna, C., Venkateshwarlu, B., Srinivas, D., Srinivas, S, D.: IoT based smart farming using cloud computing and machine learning. International journal of innovative technology and exploring Engineering, 9.3555–9.3458 (2019). https://doi.org/10.35940/ijitee.A4853.119119

21. Ahmed, N., Shakoor, N., Advancing agriculture through IoT, Big Data, and AI: A review of smart technologies enabling sustainability, Smart Agricultural Technology: 100848. ISSN. **10**, 2772–3755 (2025). https://doi.org/10.1016/j.atech.2025.100848
22. Thomopoulos, V., Tolis, F., Blounas, T.-F., Tsipianitis, D., Kavga, A.: Application of fuzzy logic and IoT in a small-scale smart greenhouse system. Smart Agricultural Technology. **8**, 100446 (2024). https://doi.org/10.1016/j.atech.2024.100446
23. Bakthavatchalam, K. et al.: IoT framework for measurement and precision agriculture: predicting the crop using machine learning algorithms. Technologies. **10**, 13 (2022). https://doi.org/10.3390/technologies10010013
24. Nawaz, M., Babar, M.: IoT and AI for smart agriculture in resource-constrained environments: challenges, opportunities and solutions. Discov. Internet Things. **5** (2025). https://doi.org/10.1007/s43926-025-00119-3

Hybrid Cryptographic Approach for File Encryption: Integrating Post-Quantum and Classical Methods

Allansius E. Pala[1], Vinodha D[1], Mary Anita E A[1], and Maria Lapina[2(✉)]

[1] Christ University, Bangalore, India
[2] North-Caucasus Federal University, Stavropol, Russia
mlapina@ncfu.ru

Abstract. Traditional cryptographic systems, such as AES and ECC, relies heavily on the computational difficulty of integer factorisation and discrete logarithm problems. But, this can be compromised by the Quantum based algorithms like Shor's. The rapid advancement of quantum computing poses a significant threat to both symmetric and asymmetric encryption. To address this challenge, a hybrid cryptographic framework is proposed by integrating classical (ECC and SES) and post-quantum (CRYSTALS-Kyber) algorithms. Quantum attacks were emulated through partial key recovery experiments inspired by Shor's algorithm, and demonstrated how selected bytes of a key could be exposed under quantum conditions. Visualisation techniques were employed to compare expected and recovered keys at the byte level, offering detailed insights into the susceptibility of classical schemes to quantum attacks. The simulation results underscore the vulnerabilities of conventional cryptography in a quantum context, while highlighting the potential of hybrid encryption as an effective transitional strategy. This study emphasises the necessity of proactive cryptographic design to ensure resilience in the post-quantum era.

Keywords: Botnet · Machine learning · Deep learning · Internet of things · Network traffic analysis

1 Introduction

With the rise of quantum computing, conventional public key cryptographic systems such as RSA and Elliptic Curve Cryptography (ECC) are becoming increasingly susceptible to attacks. Algorithms like Shor's, when run on sufficiently powerful quantum computers, can potentially break these classical encryption schemes, posing a serious threat to modern digital security infrastructures.

A hybrid cryptographic framework is proposed in this paper to resist both classical and quantum threats by combining classical cryptography (AES + ECC) with post-quantum cryptography (Kyber). The core idea is to leverage the strengths of both paradigms: the efficiency and maturity of classical methods, and the quantum resistance offered by lattice-based cryptographic algorithms.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 429–441, 2027.
https://doi.org/10.1007/978-3-032-27834-0_40

The work is structured into multiple implementation phases, starting with AES key generation and file encryption, followed by secure key encapsulation using ECC and Kyber. The encrypted keys and data are transmitted from a sender module and recovered on the receiver side. A unique aspect of this system is the simulation of quantum attacks—including Shor's algorithm to validate the system's resilience. Visual comparisons are generated to analyze how well the encrypted AES key withstands these simulated attacks. The final system demonstrates that while ECC is vulnerable to Shor's algorithm, Kyber-based encryption remains robust under the simulated quantum threat models. This hybrid encryption model offers a transitional approach for enhancing security in the emerging post-quantum era.

The rest of the paper is structured as follows. Section 2 provides the literature review on the existing methods. The methodology and implementation of the proposed work are discussed in sect. 3. Results are provided in sect. 4.

2 Existing Methods

Saiyed in [1] suggested that frameworks developed by combining classical and quantum protocols ensure more security and compatibility with existing frameworks. The authors in [2] examined the strength of Crystal-Kyber by combining it with AES. Their results show that the Kyber-based model is more secure. In [3], the authors tested a secure key exchange framework by combining classical, pre and post-quantum-based cryptography. The authors in [4] explored the strength of hybrid cryptography developed by combining post-quantum cryptography. The authors of [5, 9–12] proposed a modular blockchain framework hybridising post-quantum cryptographic signatures, quantum key distribution (QKD), and a Quantum Byzantine Fault Tolerance (Q-BFT) consensus using quantum random number generation. It provides end-to-end quantum-resistant security, eliminates vulnerabilities like Sybil, replay, and MITM attacks. The viability of executing post-quantum cryptographic (PQC) algorithms in resource-constrained IoT devices is analysed in [6]. The authors emphasise the need for lightweight, RAM-efficient solutions. NIST-validated PQC schemes are compared through practical implementations on an ESP-WROOM-32 microcontroller, highlighting their performance trade-offs and challenges in IoT environments. The triple-hybrid TLS 1.3 protocol is proposed in [7] by integrating classical, post-quantum, and quantum key distribution methods to boost security against future cryptographic threats. Experimental results show a 68% increase in communication cost, offering stronger quantum-resistant protection for secure network communications. Singamaneni in [8] proposed a novel hybrid cryptographic standard combining Curve25519 (ECC), SPECK cipher, and PHOTON hash to enhance security, speed, and energy efficiency in smart card financial transactions. The results shows that the proposed work outperforms existing RSA + AES and ECDH + AES hybrids with significantly lower computation time, with strong post-quantum resilience. The limitations of the discussed systems are presented in Table 1.

Table 1. Limitation of existing systems.

Ref no	Methodology/problem addressed	Limitation
[1]	Experimental evaluation comparing classical (RSA-4096 + AES-256), quantum (QKD + one-time pad), and hybrid (QKD + AES-256) models; metrics: encryption speed, computational overhead, network latency, and resilience against quantum attacks	Hybrid approach requires specialized quantum hardware; network latency higher than classical; cost and complexity in deployment
[2]	Quantitative analysis using datasets (NIST PQC, Google Homomorphic Encryption, Hyperledger, World Bank); methods include MCDA, AHP, MLR, ANOVA, VAR	Computational demands of PQC; challenges in legacy system integration; regulatory complexity; socio-economic impact if delayed
[3]	Design and FPGA implementation of a 3-key combiner using dual-PRF; security proofs in quantum model; performance benchmarking on FPGA	Speed bottleneck due to QKD limits; complex to combine diverse keys; newness of PQC; specific hardware demands
[4]	Literature review on PQC algorithms, NIST standardization progress; analysis of cryptographic families and hybridisation	Review article; no new experimental data; practical deployment challenges discussed but not tested
[5]	Address the vulnerability of current blockchain systems to quantum attacks, including weaknesses in cryptography, key exchange, and consensus mechanisms	Costly and challenging to implement in practical environments
[6]	Evaluated Lattice-based, Code based and Hash function-based algorithms on IoT devices using the ESP-WROOM-32 microcontroller	Limited generality, algorithms are tested on Raspberry Pi, which is relatively resourceful
[7]	Implemented a hybrid version of TLS1.3	68% increased communication overhead, impacting efficiency and practicality for real-time or resource-constrained applications
[8]	Novel hybrid cryptographic standard combining Curve25519 (ECC), SPECK cipher, and PHOTON	Relies on ECC (Curve25519), which may not offer complete resilience against future large-scale attacks using Shor's algorithm

3 Methodology and Implementation

In the proposed secure hybrid encryption system, classical cryptography (AES and ECC) is combined with post-quantum cryptography (Kyber). In the sender side, AES session key is encrypted using both ECC and Kyber, and transmitted together with the AES-encrypted files. On the receiver side, the keys are decapsulated to recover the AES key

and decrypt the data. A simulated quantum attack based on Shor's algorithm is used to evaluate system robustness, demonstrating the vulnerability of ECC and the resistance of Kyber.

3.1 Classical Cryptography

3.1.1 AES (Advanced Encryption Standard)

AES is a symmetric block cypher system that generates cypher test in fixed-size blocks of 128 bits (16 bytes). AES supports three key lengths: 16(128 bits), 24(192 bits), and 32(256 bits) bytes. In this project, AES-256 is used for simulation.

3.1.2 Elliptic Curve Cryptography (ECC)

Elliptic Curve Cryptography (ECC) is a public-key cryptographic system based on the algebraic structure of elliptic curves over finite fields. It provides high levels of security using smaller key sizes compared to RSA or other classical cryptosystems.ECC is widely used in secure communication (e.g., HTTPS, TLS, SSH, cryptocurrency wallets, etc.) due to its efficiency and strength

3.2 Quantum Cryptography

Kyber is a post-quantum cryptographic algorithm based on a hard mathematical problem called the Module Learning with Errors (MLWE) problem. Kyber-768 is a Key Encapsulation Mechanism (KEM) that is resistant to quantum attacks. It ensures security from the hardness of the lattice problem. This makes it a perfect fit for systems that aim to be secure even after quantum computers become practical. Kyber operates in the world of modular arithmetic and polynomials—unlike ECC, which uses elliptic curves.

3.3 Implementation Phases

The proposed secure hybrid encryption system is implemented in six phases.

Phase 1: AES Key Generation and File Encryption

The objective of this phase is to:

- Generate a secure AES-256 symmetric key using a cryptographically secure method.
- Split the AES key into two parts for hybrid encryption:
- One part for classical encryption using ECC.
- One part for quantum-safe encryption using Kyber.
- Use the full AES key to encrypt a set of files using the AES-CBC (Cipher Block Chaining) mode.
- Ensure the output files are secure, padded, and ready for secure decryption later using hybrid key recovery.

AES operates on fixed 16-byte blocks. If the plaintext size is not a multiple of 16 bytes, padding must be added to complete the final block. If the final block is short by r bytes. The padding scheme adds r bytes, each containing the value r. If the message

is already a multiple of 16 bytes, a full block of 16 bytes (value 0x10) is added. To ensure cryptographic security, the 256-bit AES key is generated using a cryptographically secure pseudorandom number generator (CSPRNG), and it is generated using get_random_bytes(). Initialization Vector (IV) is a 16-byte (128-bit) random value that is different for every file. This ensures that encrypting the same file multiple times produces different ciphertexts. It must be stored with the ciphertext so that decryption can be performed correctly. This ensures the key:

- Has high entropy (unpredictability)
- Cannot be guessed or reproduced
- Is suitable for long-term secure encryption

The result is a binary key that serves as the encryption key for AES-CBC. After generation, the 32-byte AES key is split into two parts (Table 2):

Table 2. After generation, the 32-byte AES key is split into two parts.

Section	Bytes	Purpose
Key part 1	Bytes 0–15	Encrypted using ECC
Key part 2	Bytes 16–31	Encrypted using Kyber-768

This split supports a hybrid encryption model where:

- One half is secured using traditional asymmetric cryptography (ECC)
- The other half is secured using a post-quantum encryption algorithm (Kyber)

This provides quantum resilience: even if ECC is compromised by a quantum attack (e.g., via Shor's Algorithm), the Kyber-protected portion still safeguards part of the key, ensuring confidentiality. Once the key is ready, each file is encrypted using AES in CBC (Cipher Block Chaining) mode.

CBC Mode prevents pattern leakage and provides strong confidentiality by chaining blocks of plaintext together. It introduces randomness using an Initialisation Vector (IV), ensuring that even identical plaintexts produce different ciphertexts.

AES in CBC mode works as follows

Inputs:

- A 256-bit AES key
- A 128-bit random Initialization Vector (IV)
- A padded plaintext file split into 128-bit (16-byte) blocks: P_1,P_2,...,P_n

Encryption Steps:

$$C_0 = IV \text{ (stored with ciphertext)}$$

$$C_1 = AES_K(P_1 \oplus IV)$$

$$C_2 = AES_K(P_2 \oplus C_1)$$

$$C_3 = AES_K(P_3 \oplus C_2)$$

$$\ldots$$

$$C_n = AES_K(P_n \oplus C_{n-1})$$

Each ciphertext block C_i depends on the previous ciphertext C_{i-1}, adding diffusion and preventing repetition. The final encrypted file includes:

- The IV (16 bytes) at the start
- The series of ciphertext blocks C_1, C_2,...,C_n

Phase 2: ECC and Kyber key Generation
The objective of Phase 2 is to generate and store the following public-private key pairs needed to encrypt the AES key in later phases

- Elliptic Curve Cryptography (ECC) key pair → for encrypting AES Key Part 1
- Kyber-768 post-quantum key pair → for encrypting AES Key Part 2

Steps to generate Elliptic Curve Cryptography (ECC) key pair
SECP384R1 is used here, a NIST-standardized elliptic curve recommended for secure applications. Here's what the name means:

- SEC: Standards for Efficient Cryptography
- P: Prime field (i.e., curve is defined over a field of prime numbers)
- 384: The bit-length of the prime field (provides ~192 bits of security)
- R1: First revision of the curve

Step 1: Select the Elliptic Curve (SECP384R1). SECP384R1 defines a specific elliptic curve with fixed parameters:

$$y^2 = x^3 + ax + b(\text{mod } p)$$

p: a large 384-bit prime number (defines the finite field)
a,b: curve coefficients (specific constants defined by NIST)
G = (xG,yG): the base point (also called generator point)
n: order of G (i.e., the number of times you can add G to itself before cycling)
h: cofactor

Step 2: Generate Private Key

- A random integer d is chosen such that: $1 \leq d \leq n-1$
- This value is generated using a cryptographically secure pseudorandom number generator (CSPRNG).
- This d becomes the ECC private key.

Step 3: Compute Public Key $Q = d.G$. Where: d: the private key (scalar), G: base point (defined by SECP384R1),$Q = (x_Q,y_Q)$: the resulting point on the elliptic curve

This operation is called scalar multiplication—it involves repeated elliptic curve point addition. The private key d is saved in a PEM file: ecc_private_key.pem. The public key Q is also saved in PEM format: ecc_public_key.pem

Steps to generate Kyber key pair

Step 1: Sample Random Matrices and Noise

The Kyber scheme generates a random matrix A with small integer coefficients. A secret vector s and an error vector e sampled from a discrete Gaussian or centered binomial distribution. These are small, noisy polynomials that make decryption work correctly while still being hard to break.

Step 2: The public key is computed using the formula:

$$t = A.s + e \bmod q$$

Where:

- A: public matrix (shared structure)
- s: secret vector (private)
- e: small error vector (adds noise)
- t: public vector = part of the public key
- q: a large modulus (a predefined prime like 3329)

This equation forms the basis of the Kyber public key.

- The security comes from the difficulty of solving for s given only A and t, even when e is small.
- This is known as the MLWE problem, which is believed to be hard even for quantum computers

Together, the public key is: $pk = (A,t)$ and the private key is: $sk = s$

After key generation, the public key is saved as kyber_public_key.bin and the private key is saved as kyber_private_key.bin

Phase 3: Hybrid Key Encryption

First AES Key Part 1 (Bytes 0–15) is encrypted using ECC public key Q_r. Then, the AES Key Part 2(16–31) is encrypted using Kyber-768. As the Kyber does not directly encrypt the data, a shared secret key is securely established between the sender and receiver and the encrypted form of AES Key Part 2 is generated as follows.

1. The sender uses receiver's public key (*pk)* and generates the ciphertext(ct) and shared secret key (K) using

$$(ct, K) = Kyber768.Encaps(pk)$$

K: shared secret key (only Alice & Bob can derive)

2. The shared key (K) is used as the symmetric key for a standard cipher AES-256 GCM to encrypt the AES Key Part 2.

$$c_data = AES_Encrypt(K, AES\ Key\ Part2)$$

Phase 4: Reconstruction of AES Key from the ciphertext
The original AES generated at phase 1 is reconstructed at the receiver end from the hybrid cypher text parts (C1, C2) by performing hybrid decryption.

1. Decapsulation at the receiver side is done as follows, using the received ciphertext(ct) and its private key(sk). This process helps the receiver to generate the same shared key K.

$$K = H\left(Decode\left(v - u^T sk (\text{mod } q)\right), ct\right)$$

Where, ct = (u,v), q = 3329, H=Hash function

2. Key part1 is decrypted as follows
   a. S = scalar_mul(ecc_sk, bytes_to_point(R))
   b. ss1 = HKDF(point_to_bytes(S) + context, length = 32)
   c. K1_enc = ss1[:32]
   d. key_part1 = AESGCM_decrypt(K1_enc, nonce1, c1, aad = context)
3. Key part 2 is decapsulated as follows

a. ss2 = kyber.decapsulate(ct, kyber_sk) # library call - > bytes
b. K2_enc = HKDF(ss2 + ct + context, length = 32)
c. key_part2 = AESGCM_decrypt(K2_enc, nonce2, c2, aad = context)

4. Reconstruction of AES key:

a. AES_key = key_part1 + key_part2

Phase 5: End to end Testing
The correctness of the reconstructed AES key is verified by comparing it with the original AES key (from Phase 1) also by ensuring that encryption and decryption were successful and no data was lost or altered.

Inputs:

1. The original AES-256 key generated in Phase 1 is taken from aes_key.bin
2. reconstructed_key.bin
3. The reconstructed AES key obtained in Phase 5 after decrypting both parts

Both files are read in binary format and compared byte-by-byte.

- If all bytes match, the decryption was successful
- If even one byte mismatches, the reconstructed key is incorrect

Phase 6: Quantum Attack Simulations
Quantum attacks (Shor's) is simulated on the system using Shor's algorithm. Shor's Algorithm Succeeds on ECC by overcoming the Elliptic Curve Discrete Logarithm Problem using Quantum Fourier Transform. But fails on Kyber as it is based on the

Module Learning with Errors (MLWE) problem, which is believed to be hard even for quantum computers. There iss no known way to reduce MLWE to a problem solvable by Shor's algorithm.

The system architecture of the hybrid cryptographic framework presented in Fig. 1. The original file is first encrypted using AES-256, generating both the encrypted file and the AES key. The AES key is then split and secured through dual key encryption one half encrypted using ECC-SECP384R1 and the other using Crystal Kyber-768, a post-quantum algorithm. These encrypted key components are transmitted alongside the encrypted file. On the receiver side, dual key decryption is performed using the corresponding ECC and Kyber private keys to reconstruct the AES key, which is then used to decrypt and retrieve the original file.

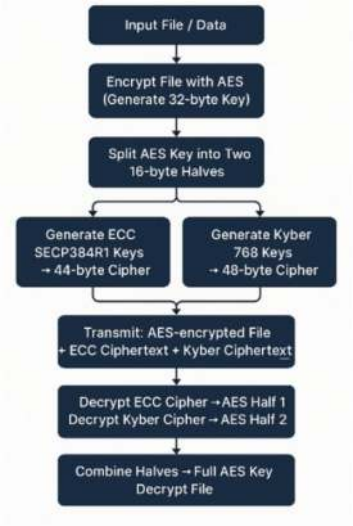


Fig. 1. System architecture.

4 Simulation Results

In this simulation, Shor's algorithm was modelled to attack the ECC private key, which has a byte size of 306 bytes. As expected, Shor's algorithm, being efficient at solving the discrete logarithm problem underlying ECC, successfully recovers the private key with minimal errors, demonstrating near-complete key compromise. The byte-wise comparison between the original and attacked keys, given in Fig. 2, shows that the bytes matched exactly. This indicates the vulnerability of ECC private keys to quantum attacks based on Shor's algorithm.

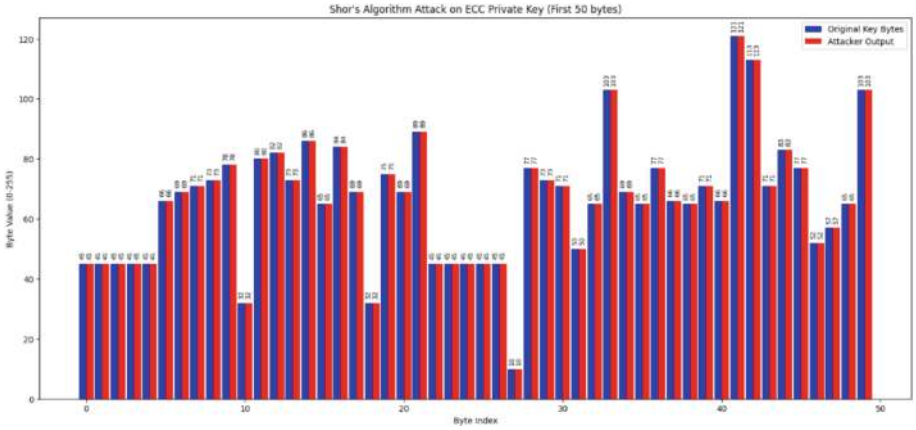


Fig. 2. Shor's attack on ECC private key.

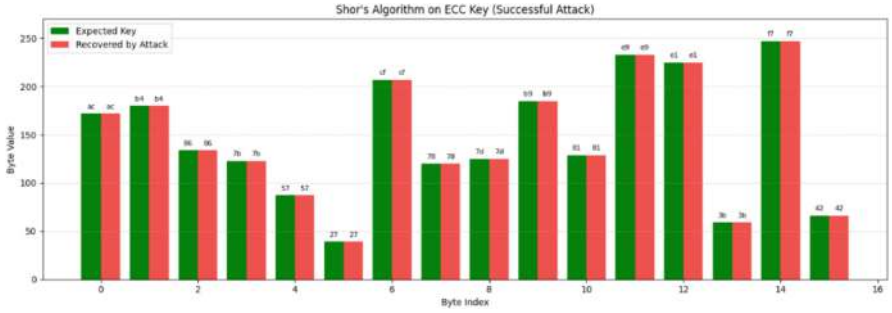


Fig. 3. Decryption of AES key part1.

In Fig. 3, the first plot demonstrates a successful quantum attack using Shor's algorithm on the 16-byte AES key half originally encrypted with ECC (Elliptic Curve Cryptography). The green bars represent the original AES key bytes, while the red bars show the bytes recovered through the simulated quantum attack. The perfect overlap between them confirms that the successful retrieval of the ECC private key using Shor's algorithm enabled full recovery of the 16-byte AES key, illustrating the vulnerability of ECC to quantum computing. This result highlights that Shor's algorithm can efficiently solve the Elliptic Curve Discrete Logarithm Problem (ECDLP), the mathematical foundation of ECC's security, thereby compromising keys secured by this method.

The attempted quantum attack using Shor's algorithm on the Kyber private key, which has a byte size of 2400 bytes is illustrated in Fig. 4. The plot compares the first 50 bytes, where the green bars represent the original Kyber key bytes and the red bars show the bytes recovered through the simulated quantum attack. The significant differences and lack of overlap between the two sets of bytes indicate the failure of Shor's algorithm to retrieve the Kyber private key. This demonstrates Kyber's robustness against quantum

attacks, as it is based on lattice problems that are currently believed to be resistant to Shor's algorithm and other quantum algorithms targeting classical cryptosystems.

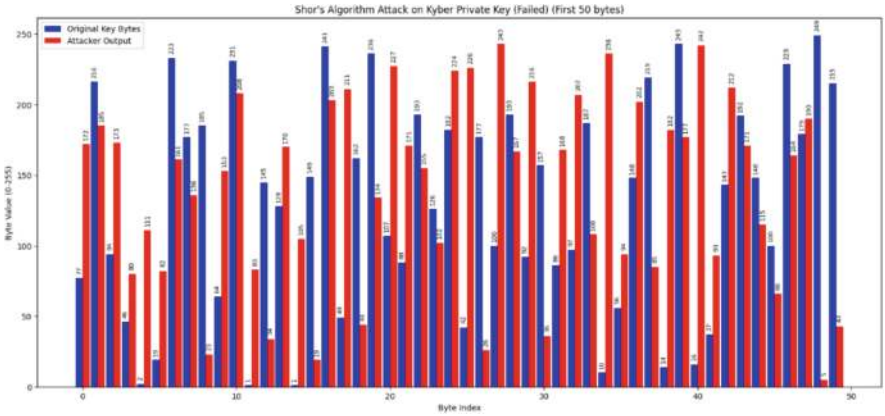


Fig. 4. Shor's algorithm attack on Kyber private key.

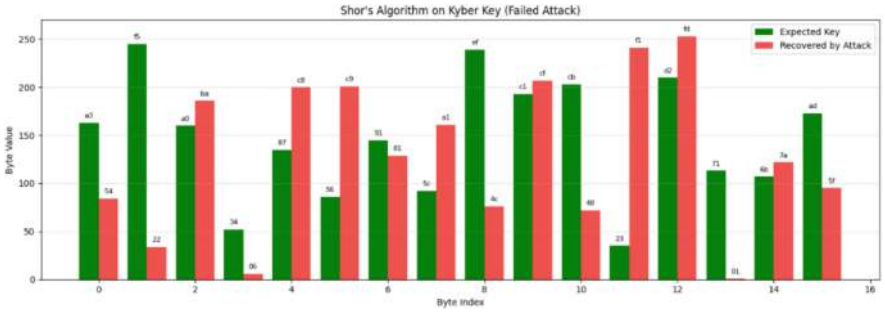


Fig. 5. Decryption of AES key part2.

An unsuccessful quantum attack using Shor's algorithm on the 16-byte AES key half encrypted using the Kyber algorithm, which is based on lattice-based cryptography, is illustrated in Fig. 5. The green bars represent the original AES key bytes, while the red bars show the bytes recovered through the simulated attack. The clear mismatch between the two, with several bytes significantly deviating, shows that Shor's algorithm fails to recover the Kyber-encrypted AES key. This reinforces Kyber's resilience against quantum attacks, since Shor's algorithm is effective only against problems like integer factorization and discrete logarithms not lattice-based problems. Therefore, Kyber remains a strong candidate for post-quantum encryption, capable of withstanding quantum computing threats.

5 Conclusion

The latest developments in the field of Bo tattack detection over the IOT and network infrastructure are deeply analysed with a summary of comparative analysis based on adopted methodologies, dataset and feature selection mechanism. The pros and cons of existing surveyed systems are tabularised to help researchers produce better technologies by addressing the highlighted challenges and focusing on reducing complexity with high accuracy by adapting better feature selection mechanisms and integrated hybrid unsupervised learning models.

References

1. Saiyed, A.I.: Hybrid quantum-classical cryptographic protocols: enhancing security in the era of quantum supremacy. Spectr. Res. **5**(1) (2025)
2. Gbadebo, M.O.: Integrating post-quantum cryptography and advanced encryption standards to safeguard sensitive financial records from emerging cyber threats. Asian J. Res. Comput. Sci. **18**(4), 1–23 (2025). https://doi.org/10.9734/ajrcos/2025/v18i4605
3. Ricci, S., Dobias, P., Malina, L., Hajny, J., Jedlicka, P.: Hybrid keys in practice: combining classical, quantum and post-quantum cryptography. IEEE Access., 1–1 (2024). https://doi.org/10.1109/ACCESS.2024.3364520
4. Cherkaoui Dekkaki, K., Tasic, I., Cano, M.-D.: Exploring post-quantum cryptography: review and directions for the transition process. Technologies. **12**, 241 (2024). https://doi.org/10.3390/technologies12120241
5. Reddy, N., Suryadevara, S., Reddy, K., Umamaheswari, R., Guttula, R., Kotoju, R.: Quantum secured blockchain framework for enhancing post quantum data security. Sci. Rep. **15** (2025). https://doi.org/10.1038/s41598-025-16315-8
6. Moura, P., Villarreal, E., Fonseca, D., Salazar, A.: Post-quantum cryptography for the internet of things: new approach. J. Eng. Exact Sci. **11**, 21741 (2025). https://doi.org/10.18540/jcecvl11iss1pp21741
7. García, C.R., Aguilera, A.C., Vegas Olmos, J.J., Monroy, I.T., Rommel, S.: Quantum-resistant TLS 1.3: a hybrid solution combining classical, quantum and post-quantum cryptography. In: IEEE Int. Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD) (2024)
8. Singamaneni, K.K.: A novel lightweight hybrid cryptographic framework for secure smart card operations. EURASIP J. on Info. Security. **2025**, 19 (2025)
9. Parfenov, D., Bolodurina, I., Lapina, M.: Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Conf. Ser.: Mater. Sci. Eng. **873**, 012020 (2020). https://doi.org/10.1088/1757-899X/873/1/012020
10. Basan, E., Makarevich, O., Lapina, M., Mecella, M.: Analysis of the impact of a GPS spoofing attack on a UAV CEUR workshop proceedings, vol. 3094, pp. 6–16 (2022). https://ceur-ws.org/Vol-3094/invited_paper.pdf
11. Proshkin, N.A., Basan, E.S., Lapina, M.A., Klepikova, A.G., Lapin, V.G.: Developing models of IoT infrastructures to identify vulnerabilities and analyse threats //. IOP Conf. Ser.: Mater. Sci. Eng. **873**(1), 012018 (2020). https://doi.org/10.1088/1757-899X/873/1/012018

12. Lapina, M., Sajid, M., Bagautdinova, A., Lapin, V., Tokmakova, M.: Comparative characteristics of the methods of using digital watermarks to protect text information. In: Raza, Z., Babenko, M., Sajid, M., Lapina, M., Zolotarev, V. (eds.) AISMA-2023: International Workshop on Advanced Information Security Management and Applications. AISMA 2023 Lecture Notes in Networks and Systems, vol. 1207. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77229-0_15

Problems and Prospects of Using Artificial Intelligence in Education

Olga Pankratova(✉), Ekaterina Konopko, and Lusine Bagdasaryan

North-Caucasus Federal University, Kulakov Avenue, 2, 355000 Stavropol, Russia
opankratova@ncfu.ru

Abstract. The article analyzes the ways of introducing end-to-end technologies and, in particular, artificial intelligence technologies into the educational process. Artificial intelligence helps to create individual learning paths, simulate the process of checking assignments by teachers, automate the creation of educational content, control, test and training tasks, analyze various aspects of the learning process, handle the results obtained, set up educational communication and implement a digital twin of the teacher. Thus, the education system is experiencing a critical stage of digital transformation associated with the introduction of artificial intelligence. However, artificial intelligence does not replace the teacher completely, but only enhances his presence, representing an additional tool for routine functions, which frees up time and makes the teacher more effective and productive for solving specific educational tasks. To date, artificial intelligence tools are not perfect, and therefore the quality of their work must necessarily be monitored by educators and specialists in this field.

Keywords: Artificial intelligence · Information technology · End-to-end technology · Educational process

1 Introduction

End-to-end digital technologies include modern promising scientific and technical areas, which currently have a significant impact on the development of most scientific and industrial sectors. End-to-end technologies include such areas as neuro-technologies and artificial intelligence, virtual and augmented reality, wireless communications, blockchain, robotics and sensor components, quantum technologies, and some other. Artificial intelligence (AI) as an end-to-end digital technology affects the development of many areas of vital activities: science, education, telecommunications, agriculture, financial sector, industry, medicine, jurisprudence, public administration, energy industry, trading, logistics, etc.

In recent years, AI has been actively introduced into the educational sphere. Leading Russian universities (HSE, MIPT, ITMO) are becoming drivers of innovation [1]. Thanks to the support of government programs and the private sector, the ROI from the introduction of AI in educational activities is 150–300% over 3–5 years [2]. It is advisable to automate routine intellectual activity for large educational projects [3] involving

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 442–451, 2027.
https://doi.org/10.1007/978-3-032-27834-0_41

a large number of students, otherwise the effectiveness of implementing AI technologies will be minimal [4]. Thus, artificial intelligence can help in the organization and implementation of the educational process in cases where there are repeatable intellectual operations.

The quality of AI work should be monitored by educators and specialists in the field of AI. At the same time, intelligent systems do not replace teachers, but help to facilitate the achievement of specific educational objectives [5]. In education, and especially for children, it is important to take into account the criteria of security, confidentiality, inclusivity, fairness, as well as to observe the principle of ethics, which at the moment is not always typical for the functioning of AI systems. The use of AI is not an end in itself, but a means to achieve the effectiveness, quality and accessibility of education. The purpose of the article is to analyze the main tasks of the education system that can be solved with the help of AI.

2 The Main Part

2.1 Creation of Individual Learning Paths and Adaptive Learning

Currently, with the help of AI, it is possible to create individual training programs taking into account the personal characteristics and needs of the student, as well as the requirements of the labor market. AI is able to form recommendations for the development of required competencies using the student's input data. Based on the formed recommendations, AI can select the content of training from an internal database and from open sources for each student [6].

In Russian education, AI is being integrated into the educational process through the creation of personalized educational paths for each student, automatic assessment and feedback systems are used, and academic risks are predicted [7]. For example, the Russian online platform for adaptive mathematics education for senior and first-year students, Plario (https://plario.ru/, Enbisys Company, Tomsk State University) takes into account the characteristics of a particular student, his level of knowledge and skills. This intelligent online learning system can vary the sequence of teaching materials in accordance with the student's individual progress and how he coped with tests and assignments. Based on the results of the entrance testing, a digital twin of the student is created, which is updated and supplemented during the training and receipt of new information. Based on the analysis of data about the student, the system offers him the amount of educational materials of varying degrees of complexity that will be sufficient to master each skill. To eliminate the probability of mechanically memorizing the correct answer, redundancy of content is provided [8].

Another Russian resource with a personification function is the resident of the Skolkovo Foundation, the Skyeng English Language Training School (https://skyeng.ru/). Among the system's capabilities are digitization of a student's speech, automatic determination of their language level, and further personalized learning based on the student's skills.

Rostec Corporation's subsidiary Roselektronika has developed a system for assessing the emotional state of students, which makes it possible to adapt and individualize learning more successfully based on the collected data. The technology consists of a

monitor with cameras and special software that analyzes the emotions on the student's face to eliminate excessive fatigue and identify problems with understanding the material. The system provides an assessment of the influence of the psychophysical state on the assimilation of educational material. So the teacher can improve the quality of his work by responding promptly to the changing environment in the classroom.

2.2 Independent Assessment of Learning Results

AI has learned to model the process of teachers checking assignments and has recently been increasingly used for automatic evaluation [9]. Intelligent systems analyze learners' responses, provide individual feedback, and are capable of making learning recommendations [10]. AI systems include technologies for text recognition and related elements, technologies for converting document images into machine-readable text, as well as speech analytics. Some platforms (for example, Gradescope, https://www.gradescope.com/) are possible to find out and analyze students' handwritten responses and give grades for their works.

One example of automated management of learning outcomes using AI is Russian educational platform and online course builder Stepik (https://stepik.org/). In order to scale education and due to the impossibility of manually checking a huge number of tasks for students, automated verification is used on the platform [11]. Assessment automatization is carried out for a wide range of tasks, including programming and data analysis, as well as for tasks of olympiads and competitions. After the evaluation, users receive instant feedback on the identified errors. The data on the verification results are used to improve the content of courses, as well as to create more personalized training programs [12]. In addition, AI can be used to assess the development of flexible skills, such as the ability to work in a team, or to diagnose communication skills [13].

Automation of the review of abstracts and term papers can occur according to the following scheme: 1) the teacher forms a task for students; 2) the student prepares and uploads the result of the task to the database of the intelligent system; 3) the intelligent system analyzes the work and forms a summary, statistics, borrowings; 4) the teacher studies the summary of the work and the identified parameters; 5) the intelligent system forms a list of control questions based on the results of the task; 6) the student answers these questions; 7) the intelligent system evaluates the correctness of the answers; 8) the teacher validates and evaluates; 9) the intelligent system forms recommendations for the development of the student.

Proctoring systems that evaluate the behavior of students during remote forms of control and eliminate cheating can also be attributed to the category of independent assessment of learning outcomes.

2.3 Automatic Creation of Training Content, Tests and Assignments

This is another task that modern intelligent systems can handle. The Russian GigaChat neural network (https://giga.chat /) allows you to create tests and exercises in various fields of knowledge in a convenient communication format. This neural network understands text messages, images, links, audio messages, files and can generate tests based on specified parameters. GigaChat creates and edits tests and assignments in one dialog.

With high-quality knowledge of the Russian language and trained on data from various scientific domains (medicine, economics, mathematics, etc.), the neural network allows you to create high-quality tasks in a few minutes, saving time and resources of the teacher.

Another similar technology is the Smartest Learning resource (https://smartest.io /), which is used to create tests, crosswords, and other practical tasks based on the analysis of uploaded educational content.

When designing training materials and tests, the following algorithm can be used: 1) the teacher describes the requirements for the material (subject, scope, target audience, goals and objectives, etc.); 2) the AI generates the structure of the educational material; 3) the teacher corrects and complements the generated structure; 4) the AI generates the content of educational materials step by step; 5) the teacher corrects, clarifies, edits the created content; 6) the AI generates control and measuring materials; 7) the teacher checks the result, correcting and supplementing it.

Currently, the content of education requires frequent adjustments and updates caused by changes in economic and socio-cultural reality. Using only open neural networks, you can reduce the time required to prepare training materials several times. As a result, the rapid development of new educational materials, updating and adapting existing materials to the needs of students is ensured. AI allows for a more flexible approach to the content of learning. Thus, AI speeds up the teacher's work, but it cannot do all the work for him. At all stages, the teacher controls the process and the result obtained with the help of AI.

The researchers believe that in the future, AI systems trained on big data arrays will be able to create accompanying content to explain the main content of learning. The AI will generate text additions or relevant visual accompaniment for a more detailed description of the topics being studied. The created content can be saved and, if necessary, returned to it for study and repetition [14]. RAG (Retrieval-Augmented Generation) technology is already capable of extracting information from available databases and generating accurate responses based on the user's request. Such systems can provide answers to students' specific questions on specific topics, such as "how to reduce fractions" and others. To get an answer, there is no need to study all the material on the topic, it is enough to set a specific request to the system.

2.4 Organization of the Educational Process Based on Educational Analytics

AI is used to analyze different aspects of education. Process Mining, which allows managing systems using digital footprint data, is becoming increasingly relevant in educational activities, as more and more processes are being logged and digitized [15]. For example, at the Moscow Electronic School, technology helps teachers predict how much time students will spend on a particular activity in the classroom or at home – completing assignments, a test, or a practical exercise. Teachers can use this forecast to plan lessons or to monitor whether the class is moving efficiently according to the curriculum.

Another example is the Russian Electronic School (https://resh.edu.ru/). The resource contains more than six thousand interactive lessons in all subjects for students from

grades 1–11. The resource has implemented algorithms for collecting analytics, in particular, searching for knowledge gaps and content filtering for the selection of materials. Based on the collected data, the system can adapt the rate and content of the training.

In educational analytics, it is possible to study the following parameters: student engagement, emotional background, speech patterns, audience workload during offline learning [16]. With the help of intelligent systems, it is possible to analyze the quality of educational materials based on calculating the percentage of correctly completed course assignments and taking into account the speed of responses. This analysis will make it possible to identify incorrectly written assignments and topics that are difficult to study in order to further adjust the forms of presentation of the material. Students' reflection data is also often used to characterize the educational process [17]. This data is easily collected, however, it may contain a certain percentage of useless information from students and require preliminary cleaning.

2.5 Automatization of Communication

For this task, educational websites and applications use chatbots capable of finding answers to repeated multiple times questions [18]. The questions can be of various kinds: from organizational to issues related to the content of individual subjects. Some chatbots can also perform the functions of automatic assessment of tasks and tests [19]. One of the most common technologies for chatbots is to search for keywords and their combinations and give a pre-recorded response. For more complex tasks (writing and translating texts, code, etc.), chatbots based on neural networks are used [20]. This category includes ChatGPT 5 from OpenAI, GigaChat from Sberb, YaGPT-5.1 from Yandex, Claude from Anthropic, DeepSeek from DeepSeek AI, etc. These language models are trained on terabytes of texts with several billion (up to several hundred billion) parameters.

As an example, let's consider an AI resource, an LLM-based university assistant from the Sber company based on GigaChat, with which applicants, students, their parents and university staff can search for the necessary knowledge and find answers to questions about studying at a university in a single window. The database of such a system stores the internal information of the university: educational and methodological materials, regulations, protocols, orders and other regulatory information. The organization's documents are uploaded to the knowledge base, classified, indexed. An intelligent assistant can, for example, explain to a student what to do if he has failed a test, or how to arrange an academic leave. The service will help the teacher arrange a business trip or plan career development. Thus, in educational activities, chatbots help employees and students receive information, answer questions, conduct tests, and support the learning process, including without the participation of a teacher.

2.6 Implementation of the Teacher's Digital Twin

The digital twin is a fairly new hybrid AI technology that allows teaching using a digital model created based on the analysis of the digital footprint of a real pedagogical system [21]. For example, the Russian development of DeepTalk (http://deeptalk.tech/ru /) allows you to interact with students through speech and text communication of the teacher's digital twin, adapting to the individual characteristics of each student. The

technology is able to present educational material in the form of lectures, clarify difficult points in the content, predict possible problems of the educational process, and ensure high motivation and involvement of students. The hybridity of the technology lies in the use of 1) AI for individualizing learning, 2) voice recognition systems, 3) chatbots, 4) online learning technologies. The teacher's digital twin technology, developed on the basis of neural networks, is constantly being updated and supplemented as it is applied. As a result, the system is able to answer almost any student's question.

In general, the considered tasks are solved using technologies that can be grouped into three categories (see Fig. 1):

- learner-oriented technologies;
- technologies focused on the educational system;
- technologies aimed at supporting the teacher's activities.

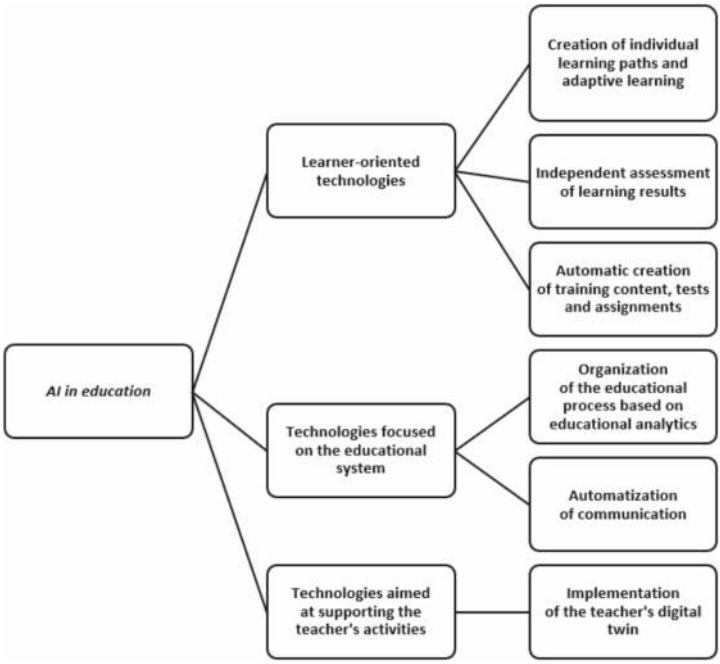


Fig. 1. Artificial intelligence in education.

From the point of view of the autonomy of the use of AI in education, three application options can be distinguished (see Fig. 2):

1. Collaborative creativity—in the process of interacting with AI, a person develops new ideas and solutions that can then be developed independently by a person or in further partnership with AI.
2. Automatization of activities—the task is divided by the user into steps, which are solved in the process of repeated application of AI with pre-prepared promts and analysis of the results after each step.

3. Implementation of AI agents—a person is involved only in setting the task to be solved and in obtaining the result of the solution, all intermediate stages are implemented using AI systems, including AI agents.

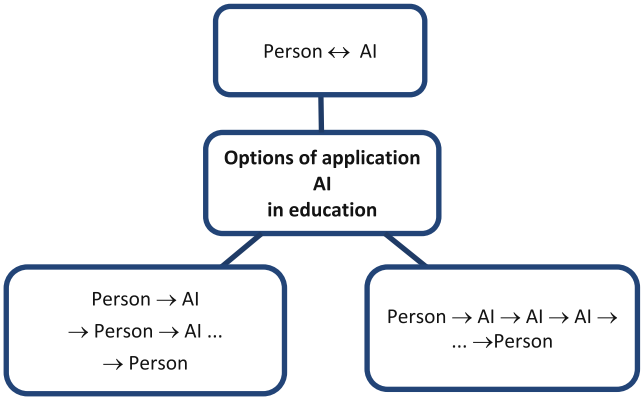


Fig. 2. Possible applications of AI in education.

Despite the advantages of using AI, there are a certain number of problems of implementing AI into education [22]. Sources of risk when working with AI can be:

1. Data. AI learns from data, and the data may not be complete enough, be of poor quality, unrepresentative, or poorly describe the process we are interested in. AI systems can be trained in the process of interacting with users, receiving data from them. In such cases, the data should be checked for validity in order to exclude incorrect facts that can lead to critical consequences when used in the field of education.
2. Prerequisites and basic assumptions. Expert judgments and basic assumptions that are inaccurate or incorrect can lead to various errors in data conversion and processing. For example, incorrect conversion of data from one unit of measurement to another can lead to inaccurate calculations and conclusions, which is unacceptable in the educational process.
3. Algorithms. The low accuracy of the algorithm, the suboptimal choice of the algorithm, as well as the uncertainty of the solutions obtained are the source of errors in AI models. For example, an error may consist in generating an image of a historical character that does not match his real appearance.

It must be remembered that AI is not capable of solving any task in areas where human presence is important. Excessive use of technology in traditionally human spheres may cause some threats:

1. Dehumanization, because technology is not able to build emotional human relationships in the full sense of the word [23].
2. Interception of control or hacking of a technical system can lead to manipulation of the system within wide limits. A person is more complex and manipulable on a smaller scale. There are other cybersecurity risks: leakage of personal data of students and

teachers, as well as documents of an educational organization, through prompts. The solution may be depersonalization of data, as well as the use of corporate products instead of public resources.
3. Copyright infringement due to the generation of content that is an object of intellectual property.
4. The risk of reducing students' self-search skills, as well as the potential weakening of critical thinking [24].
5. The rigidity of the intellectual system's activity program, while the field of education is quite flexible and is accompanied by updating educational programs, changing the composition of students and other changes [25]. AI systems may not always be ready for such changes and require adjustments of activities.
6. The need to retrain teachers to work with AI, which requires additional financial and time resources.

3 Conclusion

In conclusion, the education system is experiencing a critical stage of digital transformation associated with the introduction of AI. Key success factors include the development of high-quality and accessible AI solutions, extensive staff training, and the creation of effective mechanisms for interaction between educational organizations, government, and business. In the process of applying AI, it is important for the teacher to understand responsibility for the results obtained. All materials must be critically evaluated and checked for compliance with reality. It seems ineffective to fully trust AI at this stage of technology development. It is important to consciously choose the tasks that need to be completed with the help of AI. However, for some tasks, it may not make economic sense to use AI. Financial and time resources can multiply the economic impact of using AI. The main advantages of using AI in an educational organization are: 1) the speed of solving issues; 2) reducing the workload on employees; 3) reducing the total cost of training for both students and the educational organization; 4) ensuring the availability and attractiveness of educational materials for students; 5) the possibility of individualization and scaling of training programs.

References

1. Ranking of universities in the field of AI.: https://rating.a-ai.ru. Last accessed 02 Aug 2025
2. Measuring AI ROI.: https://www.moveworks.com/us/en/resources/blog/measuring-ai-investment-roi. Last accessed 01 Aug 2025
3. Pankratova, O., Ardeev, A., Zenkina, S., Konopko, E.: Model of organization of network project-research students activities in collaboration with city-forming enterprises. CEUR Work. Proc. **2254**, 290–296 (2018)
4. Pankratova, O.P., Abdullaev, J.A., Konopko, E.A.: The development and implementation of the online course "IT interaction in the digital economy". In: AEBMR-Advances in Economics Business and Management Research, vol. 138, pp. 171–176 (2020)
5. Kellam, H., Cortés, L.P., Gilmore, T.: Artificial intelligence as a mentor in the graduate online classroom: opportunities and challenges. Artif. Intell. Educ. **1**(1), 56–74 (2025)
6. Akyuz, Y.: Effects of intelligent tutoring systems on personalized learning. Creat. Educ., Creat. Educ. **11**(6), 953–978 (2020)

7. Parfenov, D., Zaporozhko, V., Lapina, M., Sora, D.: Development and research of algorithms for the formation the individual educational trajectories of students in the digital educational platform. In: CEUR Workshop Proceedings SLET-2019—Proceedings of the International Scientific Conference Innovative Approaches to the Application of Digital Technologies in Education and Research, pp. 258–265 (2019)
8. Bagdasaryan, L.S., Ardeev, A.H., Kulikova, T.A., Poddubnaya, N.A.: Application of digital footprint of educational intelligent technologies. Newsl. North-Cauc. Fed. Univ. **2**(101), 150–157 (2024)
9. Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., Du, Z.: Artificial intelligence in education: a systematic literature review. Expert Syst. Appl. **252**, 124167 (2024)
10. Farahani, M.S., Ghasmi, G.: Artificial intelligence in education: a comprehensive study. Forum Educ. Stud. **2**(3), 1379 (2024)
11. Abdullaev, J., Konopko, E., Pankratova, O., Taran, V., Ediev, A.: Digital education toolkit and resource review for distance interaction. Distance Educational Technologies. Collection of Works of the V International Scientific and Practical Conference. Simferopol, pp. 4–6, ID: 44059940 (2020)
12. Rugelj, J., Lapina, M.: Game design based learning of programming. In: CEUR Workshop Proceedings SLET-2019—Proceedings of the International Scientific Conference Innovative Approaches to the Application of Digital Technologies in Education and Research, pp. 29–42 (2019)
13. Gerasimova, E., Konopko, E., Usmanov, M.: Online tools for organizing students' teamwork. In: International Conference on Innovative Approaches to the Application of Digital Technologies in Education and Research, vol. 1222, pp. 167–173. Springer Nature Switzerland AG (2025)
14. Kulikova, T.A., Poddubnaya, N.A., Bagdasaryan, L.S., Ardeev, A.H.: The technique for future teachers' digital literacy development. J. Phys.: Conf. Ser. **1691**, 012129. IOP Publishing Limited (2020)
15. Kulikova, T.A., Poddubnaya, N.A., Bagdasaryan, L.S., Ardeev, A.H.: Methods for collecting and analyzing the digital footprint of project activities in the educational process. In: International Conference on Innovative Approaches to the Application of Digital Technologies in Education and Research, vol. 1222, pp. 284–294. Springer Nature Switzerland AG (2025)
16. Kavitha, K., Joshith, V.P., Rajeev, N.P.: Artificial intelligence in higher education: a bibliometric approach. Euro. J. Educ. Res. **13**(3), 1121–1137 (2024)
17. Konopko, E.A., Pankratova, O.P.: Distance learning is one of the forms of organization of university e-education. Noosphere Res. **2**, 36–43 (2021)
18. Adeyemi, I.O., Giwa, K., Giwa, F., Hassan, R., Gabriel, A.: Awareness and use of large language models among library and information science (LIS) students: implications for LIS educators. Artif. Intell. Educ. **1**(1), 42–55 (2025)
19. McCarthy, C.J., et al.: Evaluation of an artificial intelligence chatbot for delivery of IR patient education material: a comparison with societal web-site content. J. Vasc. Interv. Radiol. **34**(10), 1760–1768.e32 (2023)
20. Rangel-de Lázaro, G., Duart, J.M.: You can handle, you can teach it: systematic review on the use of extended reality and artificial intelligence technologies for online higher education. Sustainability. **15**(4), 3507 (2023)
21. Akhmadieva Roza, S., Kalmazova, N.A., Tatyana, B., Alexey, P., Molodozhnikova, N.M., Yu, S.V.: Research trends in the use of artificial intelligence in higher education. Front. Educ. **9**, 1438715 (2024)
22. Alhumaid, K., Al, N.S., Elsori, D., Mansoori, M.: The adoption of AI applications in education. Intl J. Data Netw. Sci. **7**, 457–466 (2023)
23. Konopko, E.A., Pankratova, O.P.: Professional development of teaching staff in the digital divide. Stand. Monit. Educ. **8**(3), 49–55 (2020)

24. Kulikova, T.A., Poddubnaya, N.A., Bagdasaryan, L.S.: Functional literacy formation technologies. Stand. Monit. Educ. **11**(6), 15–25 (2023)
25. Abdullaev J., Konopko E., Nersesyan E., Pankratova O.: Training of teachers for professional activity in the digital environment of the educational space. Proceedings of SLET-2019—International Scientific Conference Innovative Approaches to the Application of Digital Technologies in Education and Research, Stavropol, Dombay, Russia, pp. 205–212 (20–23 May 2019)

Assessment of Cyber Resilience of Critical Information Infrastructure to Hybrid Impacts

Yulia Parshenkova[1(✉)] and Elena Maksimova[2]

[1] RTU MIREA, 78 Vernadsky Avenue, 119454 Moscow, Russia
parshenkova@mirea.ru
[2] Institute of Management Problems of the Russian Academy of Sciences, Moscow, Russia

Abstract. Currently, the threats are hybrid in nature. Cyber attacks on critical information infrastructures can be combined with a physical incident or a campaign to destabilize incident response services. The metrics used today distinguish between cyber resilience and physical reliability [1]. The article considers an interdisciplinary approach combining the theories of crisis management and cybersecurity. The authors of the article introduce a new concept of "hybrid resilience coefficient". The paper considers the issues of developing a comprehensive model for countering hybrid threats for critical information infrastructure (CII). The article presents the results of modeling attack scenarios and their assessment, which has a cumulative destructive effect.

Keywords: Cybersecurity · Cyber resilience · Critical information infrastructure · Hybrid impact · Cyber attacks

1 Introduction

The consequences of illiterate information security management can be catastrophic: from financial losses to reputational crisis and legal sanctions [9–12]. Given this set of threats, the topic of information security comes to the fore. Hybrid threats combined with cyber attacks and physical impact, in modern geopolitical conditions, are used to achieve the greatest destructive impact on the critical information infrastructure of the state. In the article, sustainability refers to the ability of a system to withstand destructive effects, adapt to it, and restore operational functions within an acceptable time frame. CUE objects are the most attractive for an attacker in the framework of hybrid attacks [1]. Classical approaches to information security (IS) risk management are limited in nature, as they do not take into account the nature of the destructive effects that arise from hybrid attacks. The paper considers the issue of the lack of a generally accepted assessment methodology for the simultaneous occurrence of a cyber incident and physical destructive effects that reduce the resistance to IT, regarding the interaction of two separate events that are not related to each other.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 452–460, 2027.
https://doi.org/10.1007/978-3-032-27834-0_42

2 Problem Statement

The systems being developed and applied to protect significant CII facilities from malicious interference are aimed either at preventing a cyber incident or at preventing physical destructive effects. But in fact, hybrid attacks occur with greater frequency, during which either a combination of cyber incident + physical destructive effect is carried out simultaneously, or the combinations follow one another sequentially, which leads to a significant increase in damage [2].

The problem, in this case, is the lack of scientific and methodological support for assessing the risks and stability of the system relative to hybrid impacts, which can lead to an incorrect assessment of the actual level of protection of the CII, irrational operation of protection system resources, as well as inefficient system recovery after an attack [2]. The quantitative methodology considered in the article for assessing the resistance of AI to hybrid attacks would make it possible to demonstrate its real resistance, taking into account the result of the manifestation of destructive effects and an objective analysis of protection measures against hybrid threats.

3 Existing Solutions

In the issue of protecting the resistance of CII to hybrid impacts, the following solutions can be distinguished [3]:

1. Industry standards that regulate the requirements for individual CII components, but do not take into account the intersectoral increasing destructive effects.
2. Frameworks that provide, at a basic level, a methodology for assessing and managing risks, in the process of working with which each threat is considered separately.
3. Methods for analyzing the interaction of infrastructures. They identify important connections between systems, but give a superficial idea of the quantitative component of stability.
4. Dynamic modeling systems. These systems require deep knowledge and large capacities.
5. Probabilistic risk assessment models quantify impacts, but do not take into account hybrid impacts.

From the above, we can conclude that there is no methodology to assess the increasing destructive effect of hybrid attacks.

4 Discussion

A number of organizations and researchers have been involved in the development of a quantitative methodology for assessing the resistance of CII to hybrid attacks [4]. The European Union Agency for Cybersecurity (ENISA), EU, conducted a study of cross-threats to cyber-physical systems. As part of the work, a simulation was carried out of how an attack on the power grid could suspend the work of communications, healthcare and transport systems. The largest academic research center, the Lincoln Laboratory of the Massachusetts Institute of Technology (MIT Lincoln Laboratory, USA), has conducted

a number of experimental test benches for hybrid attacks. As a result of the conducted research, the low effectiveness of traditional defense systems against targeted hybrid attacks has been proven. Researchers believe that a new architecture is needed that contains elements of artificial intelligence to predict system failures with increasing destructive effects on it [4].

In order to correctly quantify the CII's resistance to hybrid attacks, we propose calculating the hybrid resistance coefficient (HRI) using the formula:

$$HRI = \left(\frac{F_0}{F_{min}}\right) \cdot \left(\frac{1}{T_{rec}} \cdot \sum Ci\right)$$

In the proposed formula:

- F_0 is the standard operating level of the system (it is assumed that initially it is 100%).
- F_{min} is the minimum level of functioning of the system achieved by it in the process of destructive impact (%).
- T_{rec} is the time taken by the system to achieve full restoration of its functionality to the level of F0 (min).
- ΣCi is the total weight of the accumulation of destructive effects, calculated on the basis of an expert assessment (a dimensionless value from 1 to n).

The higher the coefficient of hybrid cyber resilience, the more resistant the system is to hybrid impacts [5].

In the course of the work, a series of experiments was conducted in which not a real physical model was used, but a complex virtual model—a stand that simulates the behavior of a CUE object.

5 Experimental Research

To conduct an experimental study, a stand was deployed (Fig. 1), which includes:

1. Module for creating a "digital twin": software used to manage a video surveillance system, schedule management (software based on VMware ESXi)
2. Communication channel model: VoIP PBX and e-mail server.
3. The model of the energy saving system: a virtual programmable logic controller (VLC) that simulates the operation of a circuit breaker).
4. Attack Tools: A PC with software installed to generate DDoS traffic and a script simulating a damaged power cable, which in the process causes a switch to backup power with a certain time delta (Fig. 1).

Description of the presented architectural stand:

1. The "Impact System" block (red block) is a simulation of cyber attacks and physical destructive effects on the test environment. Simulates a DDoS attack on the system. Identifies weaknesses in the virtual infrastructure.
2. The "Control segment" block (blue block)—the main task is to coordinate all components of the stand and execute test scenarios. In this block, the lifecycle of virtual machines and containers is managed. The PC operator runs each stage of the hybrid attack sequentially.

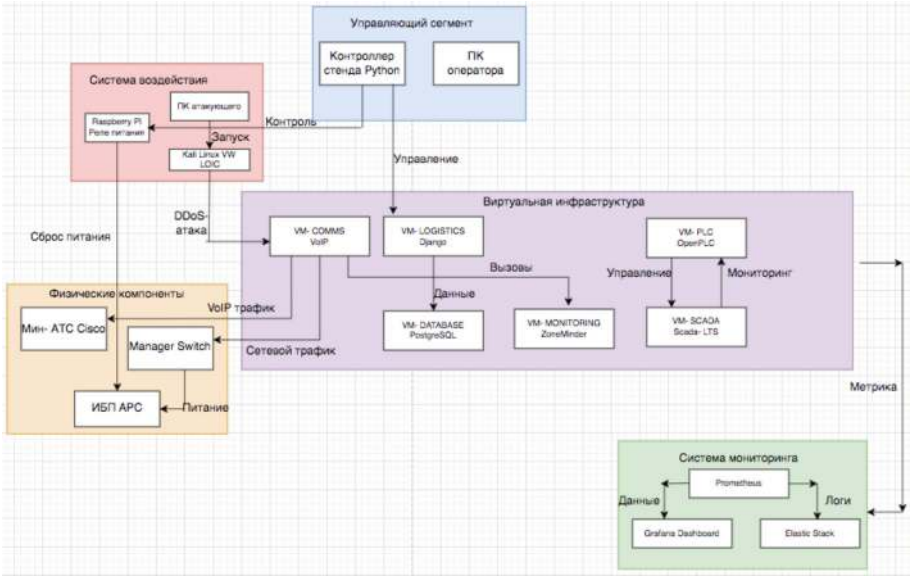


Fig. 1. Experimental stand.

3. The "Physical components" block (orange block) is the hardware used to simulate physical destructive effects. This unit is responsible for emulating power surges and power outages. The Uninterruptible Power Supply (UPS) unit simulates the operation of uninterruptible power supply systems in emergency conditions.
4. The "Virtual Infrastructure" block (purple block)—Creates a detailed model of the CII transport hub. The main advantage of the VM architecture is isolation: the failure of one VM does not affect the others and the rapid configuration change of the test environment. Also, the indisputable advantage is the ease of adding virtual machines and the ability to repeat experiments under identical conditions.
5. The "Monitoring system" block (green color block)—real-time data collection. The main goal is to collect, analyze, and visualize real-time system status data.

In the course of the work, three experiments were conducted:

1. Experiment No. 1. A separate DDoS attack on a VoIP telephony server.

The purpose of this experiment is to assess the impact of a cyberattack on the VoIP communication system of the KII.

The main initial parameters are:

Cisco PBX—corporate IP telephony;

Number of simultaneous calls: 2000;

Codec used: G.711 with a bitrate of 65 kbps per channel;

Communication channel bandwidth: 1Gbps;

The expected results of the experiment were a drop in server availability to 87–90%, the time taken to restore normal operation to 18–20 min, and the inability to make an IP telephony call in the midst of an attack.

2. Experiment No. 2. A separate "physical" impact. Damage to a power transmission line is considered as an example.

   Basic initial parameters:
   Transmission line model: 110 kV voltage, 20 km section length;
   Power consumption 85 MW;
   2 CII facilities on the site;
   UPS with an autonomous operation of 20 min, diesel generators with a start-up time of 45 s.
   The parameters of the accident were assumed to be a power line break due to a fall of a pole with a power restoration time of up to 20 min and with the shutdown of 2 districts of the city.
   Expected results:
   Complete shutdown of the main power supply for up to 20 min;
   The operating time from the backup generator is up to 15 min;
   The start time of the backup diesel generator is up to 1 min;
   Simple VoIP telephony for up to 3 min.

3. Experiment No. 3. Hybrid effect. By such an impact, we mean a previously launched DDoS attack (in 2–3 min), combined with the "damage" of the power line.

   Basic initial parameters of the DDoS attack component:
   Intensity of 60,000 packets per second;
   The bandwidth is 950 Mbps;
   Duration: 35 min;
   The number of sources is 25,000 IP addresses.
   Basic initial parameters of a physical attack:
   The power line model from Experiment No. 2;
   System recovery time: up to one hour;
   Service coordination time is up to 12 min.
   Expected results:
   Inability to start backup power due to DDoS attack;
   System recovery time is up to one hour;
   Lack of employee coordination via VoIP telephony.

6 Analysis of the Experimental Results

The results of the experimental study are presented in Table 1.

During Experiment No. 1, where the attack was directed exclusively at the VoIP telephony server, the minimum level of system functioning (F_{min}) was 45%. Below are the calculations of this value:

It is assumed that the availability to the VoIP telephony server is 0%, since the attack was successful for the attacker. The schedule management system is fully available (100%), as it uses a different communication channel during its operation. The database and power supply also have an availability rate of 100%.

Let's assume that all of the above is equivalent during the operation of the system. Then it is fair that $(0 + 100 + 100 + 100\%) / 4 = 75\%$. But we must take into account

Table 1. Comparative results of the conducted experiments.

Parameter	Experiment no 1	Experiment no 2	Experiment no 3
F_{min}, %	45	30	8
T_{rec}, min	22	18	55
Σ Ci	1,5	2	4,5
Calculation result HRI	0,067	0,093	0,0505

that the VoIP telephony server is the most important and critical for the operation of the system, therefore, it is advisable to assign it a higher weighting factor. Taking into account the weighting factor, let VoIP telephony be 40%, and the rest of the above is 20%. Then the F_{min} calculation will look like (0%* 0,4) + (100%* 0,2) + (100% *0,2) + (100% *0,2) = 45%.

The event log recorded a system recovery time (T_{rec}) of 22 min. That's how long it took for the service to start working normally after the attack was rejected.

Let us proceed to the calculation of the total weight of the increasing destructive effects (ΣCi).

For this purpose, an expert assessment was conducted on the scale of severity of the consequences of the attack. Since the gap in communication between employees did not affect important processes, but reduced their effectiveness, we can assign a weight of 1.5.

Let's proceed to calculating the hybrid stability coefficient (*HRI*) for Experiment No. 1.

Substituting the values obtained above into formula (1), we get: HRI = (100% / 45%) * (1/ (22*1,5)) = 0,067.

Based on the expected results at the beginning of the experiment, it can be concluded that the system took more than the expected time to restore normal operation (2 min more than the maximum expected time value).

Here are the calculations for Experiment No.2. During this experiment, a physical impact is implied, leading to a power outage and switching the system to a backup source.

When the main power supply was lost, the servers were forcibly rebooted. The time delta from the moment of power outage to the moment of switching to the generator was 14 min 38 s (for simplicity of calculations, we will round up to 15 min).

Thus, the availability of all services during the transition from one type of food to another is 0%. When switching to power from a backup source, the system's operating level returned to 100%. We believe that Fmin is 0% under these conditions. But taking into account the arguments outlined in the description of Experiment No. 1, we say that it is advisable to use the average values during system shutdown in the calculation. Let's assume that the total time of destructive impact is 20 min. While the shutdown of the system itself lasted no more than 15 min. Accordingly, the average level of system functioning over a time interval can be considered low. We believe that the average

indicator of the functioning of the system achieved by it in the process of destructive impact can be 30%.

The event log recorded a system recovery time (T_{rec}) of 18 min.

Let's consider the value of the total weight of the increasing destructive effects.

As in Experiment No. 1, an expert assessment is carried out. Shutdown weight of all systems = 1.5. We will take into account the loss of information in the time delta between the shutdown of the main power supply and the last automatic data saving (we will take the weight as 0.5). We get that 1.5 + 0.5 = 2.

Calculate the hybrid stability coefficient (HRI) for Experiment No.2. Substituting the values obtained above into formula (1), we get: HRI = (1,0/ 0,3) * (1/ (18* 2)) = 0,093.

In general, the expected results of the experiment coincided with those obtained.

Consider the results of Experiment No.3. It simulated a hybrid attack, during which a DDoS attack was carried out and two minutes later a physical attack (according to legend, "damage" to the power line).

It will consider the calculation of the minimum level of functioning (F_{min}) indicator.

As mentioned above, during the complex attack, the DDoS attack first disrupted communications at the facility, and then the shutdown of the main power supply led to a reboot of all its servers [6].

Under the described conditions, restoring work becomes more difficult than in Experiment No. 1 and Experiment No. 2, since the simultaneous loss of communication and basic nutrition disrupts employee interaction. Based on this, the minimum level of functioning becomes critically low and may contain only the operation of basic systems that are autonomous. This is about 8% [7].

Due to the increasing destructive effect, the system recovery time increases to 55 min. The expert assessment of the total weight of the increasing destructive effect was calculated as follows: the weight of the communication malfunction is 1.5. Shutdown of all systems = 1.5. Unplanned shutdown of systems and, as a result, data loss = 1.0. The delay in restoration work due to loss of communication is 0.5. In total, 4.5.

Let's calculate the hybrid stability coefficient (HRI) for Experiment No.3.

Substituting the values obtained above into formula (1), we get: HRI = (1,0/ 0,08) * (1/ (55* 4,5)) = 0,0505.

A visualization of the experimental results can be seen in Fig. 2.

Visualization using a histogram provides a visual representation that in Experiment No. 3 (hybrid attack), the level of functioning of the system drops many times. Accordingly, the recovery process takes longer.

Let's build a graph in Fig. 3, in which we visually present a comparison of the HRI coefficient.

Visualization of the calculated coefficients provides a visual representation that the values obtained in the cases of Experiment No. 3 were reduced by more than three times compared to the other two cases in which the attacks were isolated.

Thus, it can be unequivocally stated that the assumption is true that hybrid attacks cause more damage to CII than isolated attacks.

Moreover, in Experiment No. 3, a ripple effect was observed that was absent in the other two experiments, namely:

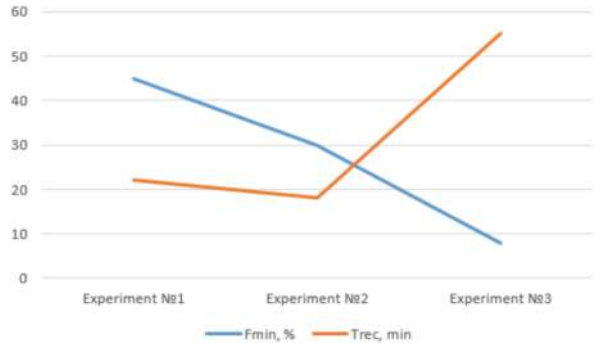


Fig. 2. Displaying the dynamics of the system's functionality.

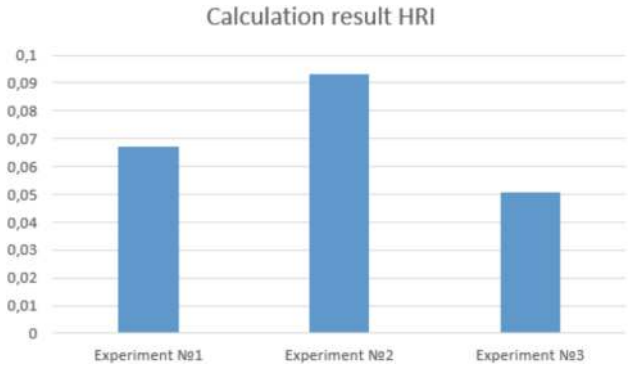


Fig. 3 Comparison of the HRI coefficient.

The first wave: communication was blocked due to an implemented DDoS attack.

Second wave: Servers were rebooted due to a power outage.

The third wave: blocked communication and disruption of interaction between employees during the implementation of manual startup of backup systems led to an increase in the T_{rec} indicator. The transaction logs were partially lost in the interval between disconnecting the main power supply and starting the backup [8].

The obtained HRI value of 0.009 during the hybrid attack turned out to be 70.5% lower than the average HRI value for Experiment No. 1 and Experiment No. 2 (0.0305). Accordingly, the damage from such an attack is more extensive.

7 Conclusion

The methodology presented in this paper makes it possible to calculate a quantitative indicator of the CII's resistance to hybrid threats. The experiments described in the article have shown that the system is ineffective if the risk of hybrid attacks is not calculated.

References

1. Maksimova, E.A.: Anthropomorphic model of states of subjects of critical information infrastructure under destructive influences. In: Maksimova, E.A., Rusakov, A.M., Lapina, M.A., Lapin, V.G. (eds.) Lecture Notes in Networks and Systems, vol. 424, pp. 569–580 (2022)
2. Maksimova, E.: Proactive modeling in the assessment of the structural functionality of the subject of critical information infrastructure. In: Maksimova, E., Sadovnikova, N. (eds.) Communications in Computer and Information Science CCIS, vol. 1448, pp. 436–448 (2021). https://doi.org/10.1007/978-3-030-87034-8_31 (дата обращения 22.09.2025)
3. Maksimova, E.A.: Predicting destructive malicious impacts on the subject of critical information infrastructure. In: Maksimova, E.A., Baranov, V.V. (eds.) Communications in Computer and Information Science CCIS, vol. 1395, pp. 88–99 (2021). https://doi.org/10.1007/978-981-16-1480-4_8
4. Maksimova, E.A.: "Smart decisions" in development of a model for protecting information of a subject of critical information infrastructure. In: Maksimova, E.A. (ed.) Lecture Notes in Networks and Systems, vol. 155, pp. 1213–1221 (2021). https://doi.org/10.1007/978-3-030-59126-7_132
5. Maksimova, E. A.: Robot technological system of analysis of cybersecurity information systems and communication networks. Maksimova, E.A., Sadovnikova, N.P., Baranov, V.V., Gromov, Y.Y., Lauta, O.S., Tret'yakova, L.V. (eds.) Journal of Physics: Conference Series, vol. 1661(1), pp. 012119 (2020). https://doi.org/10.1088/1742-6596/1661/1/012119
6. Vitenburg, E.: Decision-making based on enterprise's information system's monitoring data. In: Vitenburg, E., Nikishova, A., Maksimova, E. (eds.) 2020 International Multi-Conference on Industrial Engineering and Modern Technologies, FarEastCon 2020, p. 9271428 (2020). https://doi.org/10.1109/FarEastCon50210.2020.9271428
7. Parygin, D.: Urban areas development analysis and management decision support based on residential real estate objects evaluation methods. In: Parygin, D., Masaev, S., Malikov, V., Maksimova, E., Sadovnikova, N., Finogeev, A. (eds.) Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019, 9117312, pp. 184–189 (2020). https://doi.org/10.1109/SMART46866.2019.9117312
8. Maksimova, E.: The model of data protection in the geographic information system of housing and communal services. In: Elena, M., Yuliya, Z., Ekaterina, V., Marina, O., Lily, T., Alexey, B. (eds.) Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018, 8746933, pp.104–108 (2018). https://doi.org/10.1109/SYSMART.2018.8746933
9. Parfenov, D., Bolodurina, I., Lapina, M.: Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Conf. Ser.: Mater. Sci. Eng. **873**, 012020 (2020). https://doi.org/10.1088/1757-899X/873/1/012020
10. Maksimova, E., Lapina, M., Lapin, V.: Synthesis of Models for Ensuring Information Security of Subjects of Critical Information Infrastructure under Destructive Influences CEUR Workshop Proceedings. **3094**, 108–117 (2022) https://ceur-ws.org/Vol-3094/paper_11.pdf
11. Basan, A.S., Basan, E.S., Lapina, M.A., Kormakova, V.N., Lapin, V.G.: Security methods for a group of mobile robots according to the requirements of Russian and foreign legislation. IOP Conf. Ser.: Mater. Sci. Eng. **873**(1) (2020)
12. Tchernykh, A. et al.: Towards Mitigating Uncertainty of Data Security Breaches and Collusion in Cloud Computing, pp. 137–141 (2017). https://doi.org/10.1109/DEXA.2017.44

Network Traffic Classification Using Complex Machine Learning Methods

Tatiana Peleshenko[1], Dmitry Ardeev[1(✉)], Irina Ivanova[2], and Mohammad Sajid[3]

[1] North Caucasus Federal University 2, Kulakov Avenue, Stavropol 355017, Russia
ardeev.dima345@gmail.com
[2] MIREA — Russian Technological University, Moscow, Russia
[3] Aligarh Muslim University, Aligarh 202002, India

Abstract. This paper analyzes ensembles of machine learning methods combined with clustering methods for traffic classification problems using the CIC-IDS2017 [1] dataset. Error matrices and tables with accuracy parameters are provided. This research allows us to improve methods for detecting unfamiliar traffic in order to ensure the protection of information of various organizations.The main metrics used in the work are accuracy, integrity and correctness of predictions. The study proves the hypothesis that strategically combining clustering and ensemble methods can statistically significantly improve the accuracy of network traffic classification. Using a hybrid approach helps reduce false positives. Thus, the proposed methodology allows for the improvement of information security mechanisms through more reliable identification of threats in the network traffic of organizations. The work also identifies promising directions for further research related to the adaptation of the method to streaming data and the application of more sophisticated deep clustering algorithms.

Keywords: Clustering · Dataset · Data · Ensembles

1 Introduction

Computerization is a phenomenon that has affected the entire world in the late 20th and early 21st centuries. Technologies that allow information to be transmitted from one end of the planet to the other at speeds close to the speed of light have brought humanity to a fundamentally new level of development. Despite the fact that the first prototypes of the modern global Internet network, such as, for example, the Arpanet, had a completely military purpose and were intended for the exchange of information between various military structures, The peaceful development of network technologies has enabled human civilization to make an unprecedented leap in development in a relatively short time. From the first primitive websites, unencumbered by any web design. Just 40 years passed before email services and entire IT giants, whose domains include various internet services, emerged. With the influx of large capital into this technology sector and its rapid development, security threats arose, requiring the prompt

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 461–470, 2027.
https://doi.org/10.1007/978-3-032-27834-0_43

development of measures to prevent cybercrime. This paper examines methods for recognizing dangerous traffic based on machine learning and clustering. The CIC-IDS2017 [1] dataset was used for the study.

2 Clustering Methods

2.1 K-Means Clustering Method

The k-means clustering method is based on minimizing within-cluster variance and assumes that each cluster can be represented by its own centroid. The algorithm iteratively recalculates centers and redistributes objects, striving for configuration stability. It is the most popular data clustering algorithm [2]. This method places particular emphasis on the selection of the number of clusters, centroid initialization, and methods for assessing clustering quality. The K-means method was proposed by Stuart Lloyd in 1957 [3] and remains one of the most widely used algorithms due to its simplicity and efficiency [4]. The problem is formulated using the formula:

$$J(C) = \sum_{i=1}^{k} \sum_{x \in \mathrm{Ci}} \|x - \mu i\|^2 \tag{1}$$

Where μi is the centroid of cluster Ci, calculated as the average of all points in the cluster. The algorithm consists of the following steps: 1. Initialization of centroids: k initial centroids μ1, μ2, μk are selected; 2. Assignment of points xj refers to the closest centroid:

$$C_i = \left\{ x_j : \left\| x_j - \mu_i \right\| \leq \left\| x_j - \mu_l \right\|, \forall l \neq i \right\} \tag{2}$$

Recalculation of centroids: for each cluster a new centroid is calculated using the formula:

$$\mu_i = \frac{1}{|C_I|} \sum_{x_i \in \mathrm{C}} x_j \tag{3}$$

Stopping criteria: The algorithm terminates if: the centroids stop changing significantly, the maximum number of iterations is reached, and the value of J(C) stabilizes. The main advantages of this method are: ease of implementation and interpretation, efficiency of working with large data sets, scalability: supported by most libraries, for example, the scikit-learn library was used for this study; Robustness to noise: if the data does not contain strong outliers. The main disadvantages are: sensitivity to initial centroids, the need to pre-specify the number of clusters (k): If k is chosen incorrectly, clustering will not be performed correctly; instability to outliers: since centroids are calculated as the mean, outlier points greatly influence the result; assumes a spherical shape of clusters; does not take into account the hierarchical structure of the data: k-means produces a "flat" clustering, while some data have nested groups [5]. Thus, we can conclude that this clustering method is universal, but not ideal. Its effectiveness depends on the nature of the data and the correctness of the initial parameters. To improve the accuracy, this study combined the clustering method into ensembles with the five basic machine learning methods described above.

2.2 K-Medoids Clustering Method

This clustering method is an alternative to the classic K-means method, where centroids are the mean values of point coordinates. The key feature of this method is the use of medians, real data objects, to represent centers [6]. A medioid is selected as a point in a cluster whose sum of distances to other points in the cluster is the smallest. This principle of selecting a real point gives the K-medoids method an advantage in terms of robustness to data outliers and noise, compared to K-means [7]. The reasons for its robustness to data outliers are as follows:

1. Using a real data point. For an outlier to become a medoid, the sum of its distances to all other points in the cluster must be less than the sum of the distances from any other point in the cluster to all others. This is unlikely, since the outlier is often very far from the main mass of cluster points, and the sum of the distances from it to them will be many times greater.
2. Minimizing the sum of distances: The algorithm optimizes the sum of absolute distances or other metrics that are less sensitive to outliers than the squared distance that the K-means method minimizes. The squared distance in K-means further increases the influence of outliers on the centroid position. Minimizing the sum of direct distances in K-medoids makes it more robust.
3. The SWAP mechanism – during the optimization phase of the main algorithm implementing the K-medoids PAM method, only replacements of the current medoid with another real point within the cluster are considered. An outlier is not allowed to replace the center point, as it must produce a smaller sum of distances. If the outlier fails to become a medoid, then it remains in the cluster, but only as a peripheral point, and does not distort its center. This method can work with any distance metric, or measure of similarity and dissimilarity, because it only operates on pairwise distances between points and does not require calculating the average [8].

This makes it applicable to categorical data or data that uses complex structures (e.g., Hamming distance), where calculating the "mean" of a point is meaningless. However, it also comes with inherent computational complexity. This is especially true at the swap stage, where all possible substitutions of a medoid for a non-medoid are tested. For large datasets, this may be impractical. Various optimizations have been developed to address this issue (for example, CLARANS) [6].

3 Comparative Analysis of Methods

This study analyzed ensemble machine learning methods, which consist of combinations of data clustering and machine learning, combined to improve the accuracy of dataset measurements and network infrastructure attack predictions. Below we present the results of a machine learning model based on support vector machines [9] and random forests [10]. To train the model, the data is initially preprocessed. The goal of this model is to identify linear features that will subsequently be used to separate traffic into begin/attack features. A logistic loss function is used for training. The result of training [11] the model on the CIC-IDS2017 data network is presented in Fig. 1 as a confusion matrix.

The first machine learning ensemble is an integration of the K-means clustering method and the logistic regression model [12]. Clustering helps to identify homogeneous subsets within a sample, which improves the quality of analysis. Fig. 8 then shows the prediction error matrix obtained from training the ensemble of machine learning models on the CIC-IDS2017 dataset.

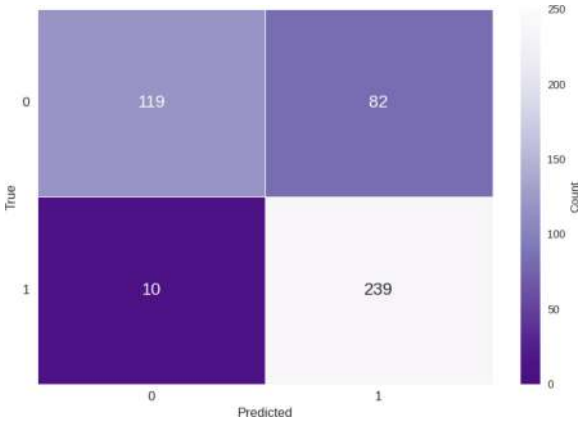


Fig. 1. Error matrix of the SVM + Kmeans ensemble.

The obtained result obtained during training on the CIC-IDS2017 dataset also shows the low efficiency of using the support vector machine [13] and k-means clustering. Despite the fact that the result of predictions of standard traffic is 119 markers, and the attacker's is 239 markers, the number of errors is still quite significant, which also calls into question The feasibility of using this ensemble of machine learning methods for traffic classification tasks is demonstrated. Further, Table 1 presents more complete data on the results of training the ensemble of methods on the dataset.

Table 1. Results of the support vector machine and K-means clustering methods.

	Precision	Recall	F1-score	Support
Begin	0.92	0.59	0.72	201
Attack	0.74	0.96	0.84	249
Accuracy			0.80	450
Macro avg	0.83	0.78	0.78	450
Weighted avg	0.82	0.80	0.79	450

The data presented allows us to compare the results of this ensemble with the previous one. Compared to the logistic regression method, the indicators increased by only 2%, making this increase negligible for a more meaningful comparison with the previous

ensemble and a standard support vector machine. This ensemble shows one of the lowest performances for solving machine learning problems for traffic classification. While this efficiency might be considered acceptable in other applications, it's still quite critical for network infrastructure. Due to the high rate of traffic prediction errors, critical local networks could be at risk. The following machine learning ensemble is a combination of K-means clustering and random forest methods. This ensemble, through a combination of clustering methods and subsequent data classification, can demonstrate high accuracy in solving the machine learning problem of detecting intrusions into a local area network. As with all previous applications of k-means clustering, this method should help to reveal the internal structure of the data and assist with further classification using the random forest method. After the clustering stage, the resulting clusters will be used either for sample segmentation or to generate additional features. In the next stage, the ensemble applies the random forest method—an algorithm based on constructing a set of decision trees trained on a random subsample of data and features. Random forest has high resistance to overfitting and copes well with imbalanced data, which can provide high classification robustness. Random forest has high resistance to overfitting and copes well with imbalanced data, which can provide high classification robustness [14], taking into account all structures. Fig. 2 shows the confusion matrix obtained when training this ensemble on the CIC-IDS2017 dataset.

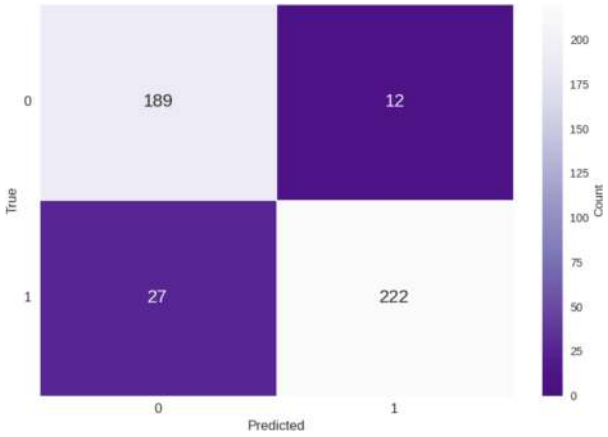


Fig. 2. Error matrix of the K-means clustering ensemble and random forest.

The confusion matrix presented above reflects the results obtained when training an ensemble of machine learning methods consisting of K-means and random forest on the CIC-IDS2017 dataset. The matrix shows the distribution of true and predicted classes by the following categories: the number of correctly classified examples of attack traffic is 189 markers, the number of incorrectly classified examples of attack traffic as allowed is 12—markers, The number of incorrectly classified data on authorized traffic as attacking is 27 markers, and the number of correctly classified data on authorized traffic is 222. Table 2 with the accuracy percentage of this ensemble is provided below.

Table 2. Results of the random forest method and the K-means clustering method.

Method	Precision	Recall	F1-score	Support
Begin	0.88	0.94	0.91	201
Attack	0.95	0.89	0.92	249
Accuracy			0.91	450
Macro avg	0.91	0.92	0.91	450
Weighted avg	0.92	0.91	0.91	450

Based on the obtained data, it can be seen that this method shows the highest accuracy among all ensembles using the K-means clustering method, which makes it the most suitable for solving machine learning problems for detecting intrusions in an local area network. Compared to the single-layer random forest method, accuracy dropped by 7%. However, if a modified dataset is used, the single-layer random forest method is highly likely to suffer performance losses and, despite its high accuracy, fail to identify the problem in time. Below, we will examine a set of basic machine learning methods in conjunction with the K-medoids clustering method. Although the core idea of this clustering method is to improve upon the basic k-means method, the solutions applied in this clustering approach differ from the original. This may impact the final data, values, and metrics obtained during the study [15]. Such changes can play a critical role in the cybersecurity of the network topology. Next, we discuss a method based on the interaction of the K-medoids method and support vector machines.

This combination is an approach that combines robustness to outliers and a clear class separationboundary. The confusion matrix is shown in Fig. 3.

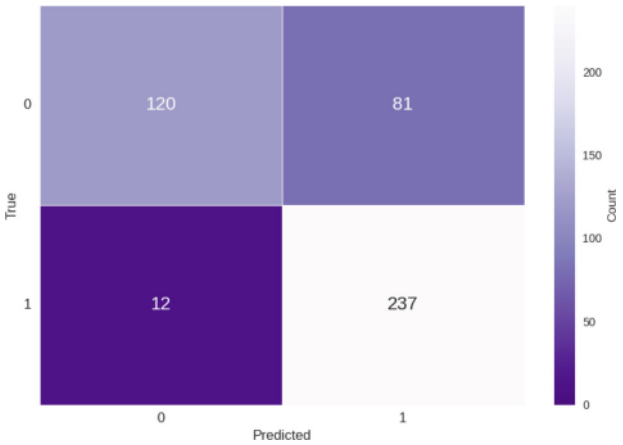


Fig. 3. Error matrix.

Attack traffic was correctly identified, namely 237 markers. Normal traffic was correctly identified, namely 120 markers. Normal traffic was incorrectly classified as attack traffic, namely 81 markers. Falsely attacking traffic was classified as normal – 12 markers. Below is a table with the percentage accuracy data for this ensemble of machine learning methods (Table 3).

Table 3. Results of support vector machine and K-medoids clustering methods.

	Precision	Recall	F1-score	Support
Begin	0.91	0.60	0.72	201
Attack	0.75	0.95	0.84	249
Accuracy			0.79	450
Macro avg	0.83	0.77	0.78	450
Weighted avg	0.82	0.79	0.78	450

The data presented in the table are extremely similar to those obtained when training with the previous method. There is an identical bias toward identifying normal traffic, making this method extremely similar to the previous one. A 1% increase is more of an error than a qualitative improvement. The final machine learning ensemble, using the K-medoid clustering method in this training, will be a combination of this method and random forests. Ensemble training occurs in two stages: in the first stage, the data is structured using k-medoids [16].

Dividing the sample into clusters. In the second stage, each of the identified clusters is independently trained using a random forest model—an ensemble of decision trees that effectively handles nonlinear dependencies and feature interactions and is highly resistant to overfitting [14].

This combination allows one to take advantage of cluster partitioning of data to extract local patterns, which significantly improves the overall accuracy and stability of classification, especially in problems with heterogeneous and high-dimensional feature spaces. This technique is particularly effective when working with noisy datasets. It can be effective in cybersecurity anomaly detection. Below is the error matrix obtained as a result of training on the ensemble dataset (Fig. 4).

Analyzing the obtained data, it is clear that the model correctly classified 225 attack traffic markers. The number of normal traffic markers was correctly identified at 188. There were 13 authorized traffic markers incorrectly classified as attack traffic. The number of negative traffic markers incorrectly classified as normal traffic was 24.Below is a table showing the ensemble accuracy as a percentage (Table 4).

The data presented in the table demonstrates the high efficiency of using random forests for detecting intrusions and anomalies in network traffic. This is achieved through coarser learning.

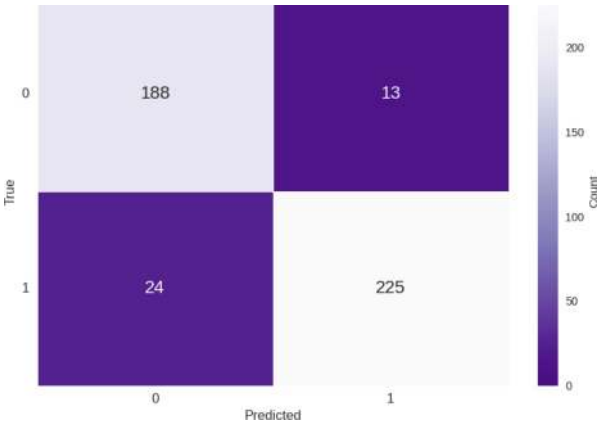


Fig. 4. Error matrix of the K-medoids ensemble and the random forest method.

Table 4. Results of the random forest method and K-medoids clustering.

	Precision	Recall	F1-score	Support
Begin	0.89	0.94	0.91	201
Attack	0.95	0.95	0.92	249
Accuracy			0.92	450
Macro avg	0.92	0.92	0.92	450
Weighted avg	0.92	0.92	0.92	450

4 Conclusion

The conducted study allows us to formulate the following conclusions regarding the effectiveness of ensemble machine learning methods integrated with clustering algorithms for solving the problem of network traffic classification. Conceptual validity of ensemble. The strategy of combining clustering methods (K-means, K-medoids) with classification algorithms has proven its conceptual validity [2]. Preliminary clustering allows for segmentation of the feature space, identification of latent structural patterns in the data, and thus creation of conditions for more efficient operation of the final classifier, especially on heterogeneous and noisy samples. Comparative analysis of model performance. Experimental results on the CIC-IDS2017 dataset revealed significant variance in the performance of different ensembles. The highest accuracy, recall, and F1-score were demonstrated by ensembles combining clustering methods (both K-means and K-medoids) with the Random Forest classifier [14]. These models achieved balanced classification accuracy of 91–92%, demonstrating successful synergy: clustering structures the data, and the ensemble of decision trees robustly processes complex nonlinear dependencies within the identified groups. Support vector machine (SVM) [17] and logistic regression-based ensembles performed significantly worse (accuracy of approximately

79 to 80%). This indicates that these algorithms in the proposed pipeline architecture either do not fully utilize the advantages of cluster analysis or are more sensitive to the specifics of the transformation of the feature space after clustering.

Comparison with base classifiers. Although the standalone Random Forest algorithm demonstrated the highest absolute accuracy (98%), the key advantage of ensemble approaches based on it is not so much the increase in this metric, but rather the potential increase in the robustness and generalization ability of the model. Dividing data into clusters can help reduce the impact of outliers and fine-tune the classifier to local features of subsets, which is critical for adapting to dynamically changing network traffic. Practical relevance for cybersecurity. From the point of view of classification problems, the most valuable ensembles are those that demonstrate high recall for the attack class while maintaining sufficient accuracy. Random Forest + K-means/K-medoids ensembles meet this requirement satisfactorily, providing robust threat detection (attack recall >0.89) with a manageable false positive rate. Directions for further research. The obtained results highlight that the effectiveness of the proposed approach depends largely on the choice of the base classifier, clustering parameters, and the integration scheme for their results. The prospects for development in this area include the study of more complex ensemble schemes (e.g., stacking), the use of other clustering algorithms, and the validation of methods on larger and more modern data sets. Thus the results of the study confirm that ensemble methods integrating clustering and classification, in particular based on the Random Forest algorithm, represent an effective and promising tool for constructing network traffic analysis systems capable of taking into account its complex internal structure.

References

1. Canadian Institute for Cybersecurity. CIC-IDS2017: Intrusion Detection Evaluation Dataset. https://www.unb.ca/cic/datasets/ids-2017.html
2. Jain, A.K.: Data clustering: 50 years beyond K-means// pattern recognition letters. (31, 8), 651–666 (2010). https://doi.org/10.1016/j.patrec.2009.09.011
3. Bock H.H. Clustering methods: a history of k-means algorithms //selected contributions in data analysis and classification 2007 161–172.
4. Ahmed, M., Seraj, R., Islam, S.M.S.: The k-means algorithm: a comprehensive survey and performance evaluation//Electronics. **9**(8), 1295 (2020)
5. Sinaga, K.P., Yang, M.S.: Unsupervised K-means clustering algorithm//IEEE access. **8**, 80716–80727 (2020)
6. Kaur, N.K., Kaur, U., Singh, D.: K-Medoid clustering algorithm-a review//Int. J. Comput. Appl. Technol. **1**(1), 42–45 (2014)
7. Harikumar, S., Surya, P.V.: K-medoid clustering for heterogeneous datasets//Procedia computer science. **70**, 226–237 (2015)
8. Shah S., Singh M. Comparison of a time efficient modified K-mean algorithm with K-mean and K-medoid algorithm//2012 international conference on communication systems and network technologies. IEEE, 2012. 435–437.
9. Cortes, C., Vapnik, V.: Support-vector networks//machine learning. **20**(3), 273–297 (1995)
10. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1995)

11. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics. **7**(2), 179–188 (1936)
12. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: Proceedings of the 2009 IEEE symposium on computational intelligence for security and defense applications, pp. 53–58 (2009)
13. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. //data mining and knowledge. Discovery. **2**(2), 121–167 (1998)
14. Breiman L. Random forests//machine learning 2001 45 1 P 5–32. https://link.springer.com/article/10.1023/A:1010933404324
15. Jordan Journal of Electrical Engineering: Clustering performance analysis of the K-Medoids algorithm for fingerprint similarity. Metr. S Jordan J. Electr. Eng. **10**(3) (2023)
16. Naalt T. K-Medoids Clustering: Global Prosperity 2023//Kaggle 2023. https://www.kaggle.com/code/tarktunataalt/k-medoids-clustering-global-prosperity-2023
17. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory (COLT '92). Association for computing machinery, pp. 144–152, New York, NY, USA (1992)
18. Jordan Journal of Electrical Engineering: Clustering performance analysis of the K-Medoids algorithm for fingerprint similarity. Metr. S Jordan J. Electr. Engineering. **10**(3) (2023)
19. Vapnik, N., Chervonenkis, A.Y.: About one class of pattern recognition learning algorithms. //Automation of solving information-logical problems/edited by Yu. Zhuravlev — M., pp. 262–273. Nauka (1963)
20. Rao, C.R.: The utilization of multiple measurements in problems of biological classification. J. R. Stat. Society. Ser. B (Methodol.). **10**(2), 159–203 (1948)

Ensemble Neural Networks for Agricultural Tasks: a Comparative Analysis of Models

Tatiana Peleshenko[1], Nikita Boikov[1(✉)], Albina Karaeva[1], Ivan Borodulin[1], and Mahfooz Alam[2]

[1] North Caucasus Federal University, 2, Kulakov Avenue, Stavropol 355017, Russia
nikitaboickov4@gmail.com

[2] Department of MCA, G. L. Bajaj Institute of Technology and Management, Greater Noida, India

Abstract. The article presents a comparative analysis of two ensemble machine learning methods: Bayesian model averaging (BMA) and Bayesian model combination (BMC) – for the task of classifying plant diseases based on leaf images from the Plant Village dataset, which contains 54,306 images divided into 38 classes. Pre-trained convolutional neural networks were used to generate base predictions, which were combined taking into account the uncertainty estimated by the Monte Carlo Dropout method. The results showed that BMC achieves the highest accuracy on the validation set (97.77%) compared to BMA (97.53%), demonstrating robustness for classes with large amounts of data. However, both methods experience difficulties with rare classes, such as Potato_healthy (recall 0.60 for BMC). The results confirm the effectiveness of ensemble approaches for automating diagnostics in agriculture and indicate the need to optimize the classification of rare classes. Future prospects include the development of data augmentation methods and the study of hybrid ensembles.

Keywords: Ensemble neural networks · Deep learning · BMA · BMC · Agriculture · Crop classification · Yield forecasting

1 Introduction

1.1 Relevance and Problem Statement

Modern agriculture must increase yields in response to population growth and resource shortages, while plant diseases cause annual losses of 20–40% [1]. Automated analysis of leaf images allows for early detection of diseases, reducing economic damage and pesticide use [2], which has a beneficial effect on precision farming and promising fog computing systems that require high reliability [3, 23]. Convolutional neural networks (CNNs) excel at classifying plant diseases by extracting visual features [4], but struggle with data imbalance and similar pathologies, which reduces accuracy for rare classes [5, 6]. Ensemble methods, including Bayesian approaches such as BMA and BMC with Monte Carlo Dropout for uncertainty estimation, improve accuracy and reliability [7, 8], although their use in agriculture, especially for rare classes, has not been sufficiently studied [9].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 471–483, 2027.
https://doi.org/10.1007/978-3-032-27834-0_44

This study compares BMA and BMC for plant disease classification using leaf images and evaluates their robustness to data imbalance. The objectives include implementing the methods, evaluating accuracy, precision, and F1 score, and proposing optimizations for rare classes. The results may improve automated diagnosis in agriculture.

2 Methodology for Studying Ensembles

2.1 Data Characteristics and Processing

The study uses the PlantVillage dataset, a benchmark dataset for plant disease classification containing 54,306 color images of 38 classes of 14 crop species (e.g., tomatoes, potatoes, grapes, apple trees, corn), including healthy plants and diseases such as bacterial spots, fungal infections, and viral diseases, with a size of 256256 pixels [9] (see Fig. 1).

Fig. 1. Sample of the displayed image.

Preliminary data processing normalized the images to 224224 pixels and applied augmentation – random rotations (0–360°), scaling (0.8–1.2), brightness shifts (±30%), and Gaussian noise ($\sigma = 0.1$) – to increase robustness, simulate real conditions, and prevent overfitting, performed in real time during training [3, 10].

2.2 Ensemble Classification Algorithms

The choice of ensemble methods and uncertainty assessment is driven not only by the task of improving accuracy, but also by the requirements for integrating models into real agricultural systems. Within modern frameworks that combine IoT, AI, and blockchain [23], the ability of a model to assess its own confidence is critical for transmitting reliable and verifiable data to distributed registries and decision support systems.

Bayesian model averaging (BMA) is a statistical approach that combines the predictions of multiple models, taking into account their posterior probabilities, which allows for uncertainty to be taken into account and improves generalization ability in classification tasks [11, 12]. The method combines model predictions using posterior probabilities to improve generalization and accuracy, which is particularly effective for plant disease classification [12, 13]. The implementation includes the following steps: data preparation using the load_data function (see Alg. 1), uncertainty assessment, combination of predictions, and evaluation of results.

Algorithm 1: Creating an image dataframe

```
Input: Dataset path
Output: Dataframe df
images ← createList();
images←createList();
labels ← createList();
labels←createList();
for i ← 0 to get_subfolder_count(dataset_path) do
    subfolder ← get_subfolder(dataset_path, i);
    for j ← 0 to get_file_count(subfolder) do
        filename ← get_filename(subfolder, j);
        image_path ← join_paths(dataset_path, subfolder, filename);
        append(images, image_path);
        append(labels, get_subfolder_name(subfolder));
df ← create_dataframe(images, labels, "image", "label");
return df;
```

Alg. 1. Data loading and systematization.

To increase the robustness of ensemble models to possible data distortions, a comprehensive approach was applied in the study [21]. It combined the analysis of neural network vulnerabilities with uncertainty assessment methods, in particular, using Monte Carlo Dropout [21].

To quantitatively assess the uncertainty of the forecast, we use the Monte Carlo method with dropout, implemented in the monte_carlo_predict function. Each model is switched to output mode with active dropout (probability 0.5) [13, 14]. This approach not only allows us to assess the confidence of the model, but also serves as a basis for identifying potentially anomalous or hostile input data, which is consistent with current research trends in deep learning security [22]. For each image, T = 5 predictions are made, which are averaged using formula (1).

$$std_m = \sqrt{\sum_{t=1}^{T} \left(p_m^{(t)} - p_m \right)^2} \tag{1}$$

It is calculated to estimate uncertainty, although in BMA it is mainly used to diagnose the quality of predictions.

This approach involves multiple predictions with activated pruning, which allows stochastic behavior to be simulated. The algorithm is described by the following pseudocode (see Alg. 2):

Algorithm 2: Monte Carlo Prediction for Uncertainty Estimation
Input: Model M, data generator G, number of samples N, iterations T
Output: Mean μ, standard deviation σ
all_preds ← createList();
for t ← 0 **to** T **do**
  reset_generator(G);
  preds ← model_generator.call(N/batch_size);
  append(all_preds, slice(preds, 0, N));
μ ← *compute_mean(all_preds_array, T);*
σ ← *compute_std(all_preds_array, μ, T);*
return (μ, σ);

Alg. 2. Monte Carlo Dropout.

For each model, 5 passes are performed (T = 5), which provides an adequate balance between the accuracy of uncertainty estimation and computational resources [14]. The aggregation process begins with calculating logits for the average forecasts of each model. Logits are calculated as $\log(p_m + 10^{-8})$, where p_m is the average forecast of model m, and 10^{-8} is added to avoid log(0). For BMA, logits are implemented using:

$$bma_{logits} = \frac{logits_1 + logits_3}{T} \tag{2}$$

From this, we find the probability using the softmax principle:

$$p_{bma} = \frac{\exp(bma_{logits})}{\sum \exp(bma_{logits})} \tag{3}$$

The Combine Predictions BMA function is an algorithm that implements the combination of predictions using Bayesian model averaging (BMA). This procedure takes as input two arrays of mean predictions (val_mean1 and val_mean3) corresponding to the first and third models, as well as a parameter T, which determines the number of iterations or samples used in the averaging process (see Alg. 3).

Algorithm 3: Bayesian averaging of a model for classifying plant diseases
Input: Average val_mean validation values, number of iterations T
Output: Predictions of BMA bma_preds, probabilities of BMA bma_probs
$logits \leftarrow log\ (val_{mean} + 10^{-8})$;
$bma_logits \leftarrow \frac{1 - logits}{T}$;
$bma_probs \leftarrow \frac{exp\ (bma_{logits})}{\sum exp\ (bma_{logits})}$;
$bma_preds \leftarrow arg_max(bma_probs)$;
return *bma_preds*;

Alg. 3. Combining forecasts using Bayesian model averaging.

The algorithm calculates the logarithms of the val_mean1 and val_mean3 arrays with a constant of 10^{-8} to prevent zero errors, storing the results in l1 and l2. It then sums l1 and l2, divides by T for averaged logits (bma_logits), applies an exponential function, normalizes by sum to form bma_probs (softmax), and sets bma_preds to the indices of the maximum values of bma_probs, returning a pointer to bma_preds.

Bayesian model combination (BMC) improves ensemble learning by integrating model predictions and uncertainties, unlike BMA, which uses posterior probabilities, by optimizing weights with uncertainty in mind [15, 16]. BMC uses the Monte Carlo method for uncertainty and the Metropolis-Hastings method with Dirichlet distribution for balanced weight optimization.

The BMC process includes data preparation, uncertainty estimation using the Monte Carlo method, weight optimization using the Metropolis-Hastings and Dirichlet methods, and final aggregation of predictions [17].

$$\overline{p_i} = \frac{1}{T}\sum_{t=1}^{T} p_i^t(c) \tag{4}$$

Where $p_i^t(c)$ is the probability of class c predicted by model i on pass t.

The method performs 5 passes (T = 5) with the dropout function enabled to improve estimation accuracy (see Alg. 4):

Algorithm 4: Forecasting using the Monte Carlo method to assess uncertainty

Input: Model M, data generator G, number of samples N, iterations T
Output: Forecasts ŷ, uncertainty u

```
reset_generator(G);
predictions ← createList();
num_steps ← N/BATCH_SIZE;
for t ← 1 to T do
    preds ← createList();
    foreach batch ∈ generator(training=True) do
        pred ← M(batch, training=True);
        append(batch_preds, pred);
    append(predictions, concatenate(batch_preds[:N]));
predictions ← stack(predictions, axis=0);
mean_pred ← predictions.mean(axis=0);
uncertainty ← predictions.std(axis=0);
return softmax(mean_pred), uncertainty;
```

Alg. 4. Monte Carlo Dropout.

The weight coefficients of the models are determined as a posteriori distribution selected from the Dirichlet distribution [17, 18]. The Dirichlet distribution density for K models with weights$W = (w_1, w_2 \ldots w_k)$, where wi ≥ 0 and $\sum_{i=1}^{K} w_i = 1$, is given by the formula:

$$Dir(w|a) = \frac{\Gamma\left(\sum_{i=1}^{K} a_i\right)}{\prod_{i=1}^{K}\Gamma(a_i)} \prod_{i=1}^{K} {w_i}^{a_i - 1} \tag{5}$$

Where,

$W = (w_1, w_2 \ldots w_k)$ – Weight vector (wi $\geq$ 0),

$\alpha = (\alpha_1, \alpha_2 \ldots \alpha_k)$ – concentration parameters (αi $>$ 0),

G() is the gamma function that provides normalization,

$\prod_{i=1}^{K} w_i^{a_i - 1}$ determines the distribution shape, where ai determines the "weight" for wi.

Weight optimization is performed using the Metropolis-Hastings algorithm, implemented in the bmc_predict_full function as follows (see Alg. 5):

Algorithm 5: A combination of Bayesian models with uncertainty

Input: List of predictions P, true labels y, uncertainties U, number of models M

Output: BMC bmc_probs probabilities

```
M ← len(P);
prior_alpha ← estimate_prior_alpha(P, y, U);
weight_samples ← Dirichlet(prior_alpha, N_samples);
bmc_probs ← zeros_like(P[0]);
total_weight ← 0.0;
for i ← 1 to M do
    log_likelihood ← compute_log_likelihood(P[i], y);
    post_weight ← exp(log_likelihood);
    if not isfinite(post_weight) then
        return Error: 'BMC weight computation failed';
    bmc_probs ← bmc_probs + post_weight * P[i];
    total_weight ← total_weight + post_weight;
bmc_probs ← bmc_probs / total_weight;
return softmax(bmc_probs, axis=-1);
```

Alg. 5. Optimizing weights using an algorithm.

The final prediction for class c is calculated as the weighted sum of the model predictions:

$$p_{bmc}(c) = w_1 p_1(c) + w_2 p_2(c) \tag{6}$$

Where $p_1(c)$ is the probability of class c predicted by model i.

During Bayesian model combination (BMC), the algorithm loads data using the load_data function, then splits it into samples and creates normalized generators. Uncertainty is estimated using the Monte Carlo Dropout method in the mc_dropout_predict function with 5 passes, calculating average probabilities and standard deviation. Weight optimization is performed using the Metropolis-Hastings method with Dirichlet distribution in the bmc_predict_full function, which determines the concentration parameters and combines the predictions.

2.3 CNN Models

The study uses two pre-trained models: "Classification of plant diseases using CNN" and "Efficient-Net-B0: detection of plant diseases." The first uses a classic CNN structure with convolutional and pooling layers for feature extraction and fully connected

layers with softmax for classification of 38 disease classes (see Fig. 2), pre-trained on PlantVillage by transferring training from ImageNet [20].

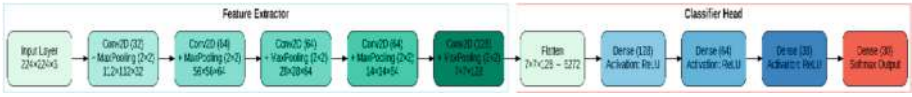


Fig. 2. Model Architecture.

The architecture of the model is described by the convolution equation:

$$(I * K)_{i,j} = \sum_{m} \sum_{n} I(i+m, j+n) \cdot K(m,n) \tag{7}$$

Where I is the input tensor (image), K is the k k convolution kernel (here k = 3), (i, j) is the position in the output tensor [19, 20]. The model is pre-trained on the PlantVillage dataset using transfer learning, which allows characteristics extracted from more general datasets, such as ImageNet, to be adapted to specific disease classes.

The EfficientNet-B0: Plant Disease Detection model is based on the EfficientNet architecture, optimized to improve computational efficiency by balancing network depth, width, and resolution [13]. A key element is the MBConv block, which uses the Swish activation function and is adapted to classification tasks by retraining on the PlantVillage dataset (see Fig. 3).

Fig. 3. Model architecture.

It uses pre-trained weights adapted to PlantVillage classification tasks through retraining. The main element of the architecture is the MBConv block, described as:

$$Y = \sigma(W * X + b) \tag{8}$$

(8)Where () = · sigmoid () – Swish, W – convolution kernel, X – input data.

Deep 3x3 convolution is applied separately to each channel to extract spatial characteristics with minimal computational load [13].

$$Y_i = W_i + X_i \tag{9}$$

Where *i* is the channel index.

3 Ensemble Study Results

3.1 Ensemble Study Results Using Bayesian Model Averaging (BMA)

As part of the study, two pre-trained models and their combination were evaluated using the BMA method on a validation sample. The results showed that the ensemble approach provided a 0.88% increase in accuracy compared to the most effective individual model (model 2), indicating a synergistic effect from combining the predictions of the two models [13, 22]. Thanks to the use of ensembles, CNN models demonstrated a significant improvement in image classification, confirming the effectiveness of this approach in plant disease recognition tasks (see Fig. 4).

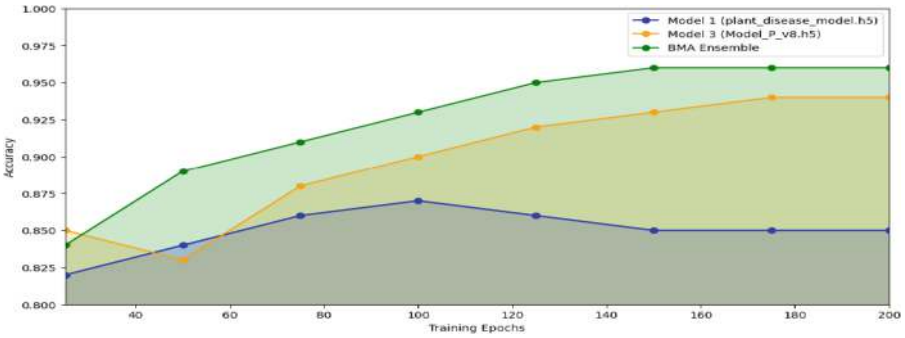


Fig. 4. Dynamics of the accuracy of the BMA model and ensemble.

Model 1 outperforms many classes in terms of accuracy, but struggles with Potato_healthy (15 examples), demonstrating lower recall and F1-score. Model 2 improves performance across most classes but struggles with limited samples. BMA improves the F1-score and accuracy by combining both models, specifically improving Apple_Apple_scab and correcting the high error rates of Model 1.

As a result, the BMA method increased accuracy by 0.88% compared to the individual models, demonstrating stability and the ability to combine their advantages. Despite a decrease in accuracy for some rare classes, this approach showed promise for further optimization of ensemble parameters and expansion of the training sample.

3.2 Results of the Bayesian Model Combination (BMC) Ensemble

The study evaluated two pre-trained models and their BMC combination with weight optimization based on prediction uncertainty. The BMC ensemble achieved an accuracy of 97.77%, surpassing the classification of plant diseases using CNN (91.75%) and EfficientNet-B0 (96.57%). The ensembles improved on CNN classification, especially for classes such as Cherry_(including_sour)_healthy (86/86), Corn Common_rust (118/119), Corn_healthy (116/116), and Strawberry_Leaf_scorch (111/111), with diagonal matrix values close to the total number of examples (see Fig. 5).

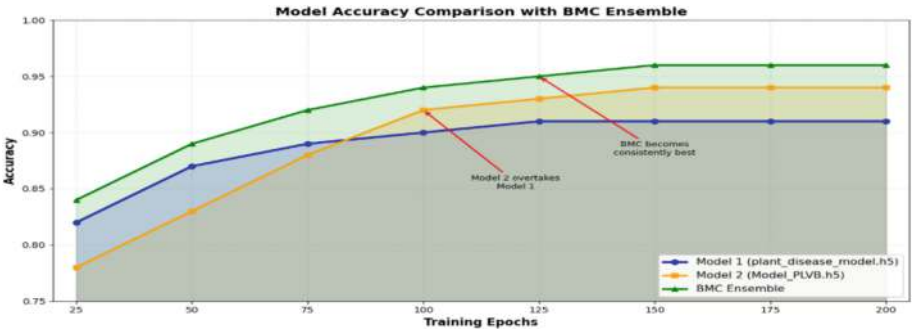


Fig. 5. Dynamics of accuracy of the BMC model and ensemble.

The BMC method effectively combines the advantages of both models, improving metrics for complex classes. For example, for Tomato_Late_blight, the F1 score reaches 0.9524 compared to 0.7676 (model 1) and 0.9514 (model 3), and for the Orange_Haunglongbing (Citrus_greening) class with 551 examples, the F1 score is 0.9973, outperforming model 1 (0.9873) and model 3 (0.9955).

Classification of plant diseases using CNN has difficulty with accuracy, especially for Potato_healthy (15 examples), correctly classifying 9 (repeatability 0.6000, F1 score 0.6429), reflecting the problems of training with small samples. EfficientNet-B0 generally performs better but struggles with complex classes such as Corn_(maize)_Cercospora_leaf_spot_Gray_leaf_spot (repeatability 0.7692).

3.3 Comparative Analysis of Ensembles

The comparative analysis evaluates Bayesian model averaging (BMA) and Bayesian model combination (BMC) based on accuracy, F1 score (macro and weighted), cross-validation stability, and performance for rare classes [13, 19, 20]. The results presented in Table 1 evaluate both methods on the PlantVillage dataset (54,306 images, 38 classes), and Table 2 highlights the improvements with BMA and BMC ensembles.

Table 1. Comparative analysis of implemented ensembles.

Criterion	Bayesian model averaging (BMA)	Bayesian model combination (BMC)
Recall (macro avg)	~0.95–0.96	~0.94–0.95
F1 (macro avg)	~0.96–0.97	~0.96–0.97
Precision (weighted avg)	~0.97–0.98	~0.97–0.98
Recall (weighted avg)	~0.97–0.98	~0.97–0.98
F1 (weighted avg)	~0.97–0.98	~0.97–0.98
Stability (cross-validation)	~0.97 ± 0.001	~0.97 ± 0.001
Ensembling Method	Probability averaging with MC Dropout	Bayesian averaging, Dirichlet distribution
Implementation Details	- MC Dropout (T = 5) - Temperature scaling (T = 1.5)	- MC Dropout (T = 5) - Dirichlet weight distribution
Strengths	- Uncertainty estimation - Improved F1 for complex classes	- High accuracy on large classes
Weaknesses	- Reduced recall for rare classes	- Low recall for rare classes
Performance on Rare Classes	Reduced recall (e.g., “Potato healthy”)	Low recall for rare classes (recall = 0.60)
Training Time	~48 min (high computational cost)	~57 min (most resource-intensive)

Stability analysis shows high consistency (~0.97 ± 0.001 when cross-checked) for all methods, although computation times vary: BMA (~48 min) versus BMC (~57 min) due to algorithm complexity. BMA uses Dropout uncertainty, while BMC integrates uncertainty with adaptive weights. All methods improve accuracy compared to single models, but rare classes require optimization with additional data or adjusted parameters.

Table 2. Improved classes using ensembles.

Improved classes	Bayesian model averaging (BMA)	Bayesian model combination (BMC)
Cherry (incl. sour) – healthy (86/86)	Accuracy: 99.8%, Recall: 1.0, F1: 0.99	Accuracy: 100%, Recall: 1.0, F1: 1.0
Corn (maize) – Common rust (118/119)	Accuracy: 99.2%, Recall: 0.99, F1: 0.99	Accuracy: 99.6%, Recall: 0.99, F1: 0.99
Potato – healthy (13/15)	Accuracy: 83.0%, Recall: 0.83, F1: 0.83	Accuracy: 84.0%, Recall: 0.84, F1: 0.84
Tomato – healthy (159/159)	Accuracy: 99.9%, Recall: 1.0, F1: 0.99*	Accuracy: 100%, Recall: 1.0, F1: 1.0
Orange – Huanglongbing (321/321)	Accuracy: 99.9%, Recall: 1.0, F1: 0.99	Accuracy: 100%, Recall: 1.0, F1: 1.0
Tomato – Septoria leaf spot (331/341)	Accuracy: 89.9%, Recall: 0.9, F1: 0.89	Accuracy: 90.0%, Recall: 0.9, F1: 0.9
Tomato – Leaf Curl Virus (298/298)	Accuracy: 99.9%, Recall: 1.0, F1: 0.99	Accuracy: 100%, Recall: 1.0, F1: 1.0
Tomato – Spider mites (204/234)	Accuracy: 87.7%, Recall: 0.8, F1: 0.87	Accuracy: 91.0%, Recall: 0.9, F1: 0.9

4 Conclusion

The study compared two ensemble machine learning methods, Bayesian model averaging (BMA) and Bayesian model combination (BMC), for classifying plant diseases using leaf images from the PlantVillage dataset. Both methods outperformed individual models in terms of accuracy, with BMC achieving the highest validation accuracy of 97.77%, slightly exceeding BMA's 97.53%. This advantage is due to the flexible combination of BMC predictions using a Dirichlet distribution to optimize weights. However, both methods struggled with rare classes such as Potato_healthy, where BMC's recall was only 0.60. In terms of computation, BMA was faster (48 min vs. 57 min for BMC) due to the additional weight optimization steps in BMC.

These results highlight the promise of Bayesian ensembles for automated diagnostics in agriculture, while also pointing to the need for improved classification of rare classes and reduced computational requirements. Future research may focus on exploring targeted data augmentation, alternative uncertainty estimation methods, and hybrid approaches combining BMA, BMC, and other ensembles.

References

1. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Front. Plant Sci. **7**, 1419 (2016)
2. Ferentinos, K.P.: Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. **145**, 311–318 (2018)

3. Shiriaev, E., Ermakova, T., Bezuglova, E., Lapina, M.A., Babenko, M.: Reliablity and security for Fog computing systems. Information. **15**(6), 317 (2024)
4. Barbedo, J.G.A.: Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 6105–6114 (2019)
6. Brahimi, M., Boukhalfa, K., Moussaoui, A.: Deep learning for tomato diseases: classification and symptoms visualization. Appl. Artif. Intell. **31**(4), 299–315 (2017)
7. Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems, pp. 1–15. Springer, Berlin, Heidelberg (2000)
8. Wilson, A.G., Izmailov, P.: Bayesian deep learning and a probabilistic perspective of generalization. Adv. Neural Inf. Process. Syst. (NeurIPS). **33**, 4697–4708 (2020)
9. PlantVillage Dataset: dataset of diseased plant leaf images and corresponding labels. [https://www.kaggle.com/datasets/abdallahalidev/plantvillage_dataset/code?datasetId=277323&sortBy=voteCount]
10. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML), pp. 1050–1059 (2016). https://proceedings.mlr.press/v48/gal16.html
11. Monteith, K., Carroll, J.L., Seppi, K., Martinez, T.: Turning Bayesian model averaging into Bayesian model combination. In: Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN), pp. 2657–2663 (2011)
12. Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., Stefanovic, D.: Deep neural networks based recognition of plant diseases by leaf image classification. Comput. Intell. Neurosci., 3289801 (2016)
13. Hoeting, J.A., Raftery, A.E.: Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. **91**(433), 179–191 (1996)
14. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T.: Bayesian model averaging: a tutorial. Stat. Sci. **14**(4), 382–417 (1999)
15. Lu, J., Hu, J., Zhao, G., Mei, F., Zhang, C.: An in-field automatic wheat disease diagnosis system. Comput. Electron. Agric. **142**, 369–379 (2017)
16. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC (2013)
17. Monteith, K., Carroll, J.L., Seppi, K., Martinez, T.: Bayesian model combination: an alternative to Bayesian model averaging. IEEE Trans. Neural Netw. Learn. Syst. **24**(6), 937–949 (2013)
18. Plant Disease Prediction with CNN Deep Learning. [https://www.kaggle.com/code/pahirathannithilan/plant-disease-prediction-with-cnn-deep-learning]
19. Plant Diseases Classification using CNN. [https://www.kaggle.com/code/pahirathannithilan/plant-disease-prediction-with-cnn-deep-learning]
20. Fragoso T. M, Louzada Neto F.: Bayesian model averaging: a systematic review and conceptual.
21. Basan, E. et al.: UAV security analysis framework. In: Kumar, S.K.B., Kim, J.H., Bansal, J.C. (eds.) Fourth Congress on Intelligent Systems. CIS 2023. Lecture Notes in Networks and Systems, vol. 865, pp. 379–391. Springer, Singapore (2024)

22. Lapina, M., Dudun, G., Kotlyarov, D., Rjevskaya, N., Subramanian, S.J.: Analysis of an existing method for detecting adversarial attacks on deep neural networks. In: Alikhanov, A., Tchernykh, A., Babenko, M., Samoylenko, I. (eds.) Current Problems of Applied Mathematics and Computer Systems. CPAMCS 2023. Lecture Notes in Networks and Systems, vol. 1044, pp. 316–329. Springer, Cham (2024)
23. Vinodha, D., Buvana, M.J., Rajalakshmi, S., ... Anita, E.A.M., Lapina, M., Framework based on IoT, AI, and blockchain for smart access to government agricultural schemes, the future of agriculture: IoT, AI and Blockchain technology for sustainable farming, 2024, pp. 142–163

Application of Machine Learning Algorithms in Recommendation Systems

Tatiana Peleshenko, Mira Ermolovich(✉), Albina Karaeva, Anton Afanasev, and Alexander Gish

North Caucasus Federal University, 1 Pushkin St., Stavropol 355017, Russia
miradudova@yandex.ru

Abstract. The article discusses the application of machine learning algorithms in recommendation systems using the example of popular platforms RuTube and Yandex.Music. The main focus is on approaches to content personalisation that improve user experience and increase audience engagement. Among these approaches, collaborative filtering stands out, based on analysing user interactions and identifying similar preferences, content filtering, which takes into account the characteristics of videos and music tracks, as well as hybrid methods that combine both approaches to improve the accuracy of recommendations. A comparative analysis of the methods used for video and music content recommendations is conducted, taking into account the characteristics of each platform. It also examines in detail the scalability issues that arise as the number of users and the volume of content increase, as well as the so-called 'cold start' effect, when there is insufficient data to accurately predict the interests of new users or recently uploaded content. Particular attention is paid to modern algorithms, including neural network models and deep learning methods, which improve the quality of recommendations, identify hidden patterns in user behaviour, and provide more accurate personalisation. The use of such algorithms contributes to increased user satisfaction, longer interaction time with the platform, and the formation of long-term audience loyalty.

Keywords: Recommendation systems · Machine learning · Collaborative filtering · Content filtering · RuTube · Yandex · Music · Personalisation · Neural network models

1 Introduction

In today's digital space, recommendation systems play a key role in the functioning of online services, helping users navigate the ever-growing volume of information and content. They analyse user behaviour, preferences, interaction history with the platform, and the characteristics of the objects themselves to generate personalised recommendations. Such systems are used to select videos, music tracks, articles, goods, and other types of digital content. Personalised recommendations allow users to spend less time searching for interesting material, increasing audience engagement and satisfaction. For the platforms themselves, effective recommendation mechanisms contribute to an increase in user return rates, longer interaction times with the service, and stronger user loyalty.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 484–493, 2027.
https://doi.org/10.1007/978-3-032-27834-0_45

The purpose of this article is to analyse the application of machine learning algorithms in the construction of recommendation systems using the example of two popular Russian online services: the video hosting service RuTube and the music streaming service Yandex.Music. The paper examines the theoretical foundations of recommendation systems, methods of collaborative, content-based, and hybrid filtering, as well as architectural solutions used for content personalisation. Particular attention is paid to comparing approaches to processing video and music content, which allows us to identify the impact of data type on the accuracy and quality of recommendations.

2 Theoretical Foundations of Recommendation Systems

Modern recommendation systems are a complex synthesis of machine learning and data analysis methods aimed at solving the problem of information overload through content personalisation. Their evolution has led to the formation of several key paradigms, each of which addresses fundamental aspects of predicting user preferences.

Historically, the core of most systems has been collaborative filtering [4], which relies on the collective wisdom of users without requiring analysis of the objects themselves. This approach is revealed through two complementary principles. The user-based method [14], focused on finding similarities between user profiles, allows us to identify hidden patterns in their behaviour, forming recommendations based on the principle that 'similar people prefer similar things.' In turn, the item-based method [14] shifts the focus to the internal connections between objects, building recommendations based on their semantic proximity, which ensures high stability and explainability, especially for platforms with a large catalogue, such as video services and music streaming.

However, the limitations of collaborative filtering, most acutely manifested in the 'cold start' problem for new users and objects, have stimulated the development of content filtering. This method offers a fundamentally different solution, relying on the internal attributes of the content. For videos, this could be genre and metadata, and for music compositions, low-level audio features such as tonality, tempo, and timbre. This approach allows for effective recommendations of new releases and niche content, but creates the risk of forming a so-called 'filter bubble,' excessively narrowing the diversity of results.

However, the limitations of collaborative filtering, most acutely manifested in the 'cold start' problem for new users and items, stimulated the development of content filtering. This method offers a fundamentally different solution, relying on the internal attributes of content. For videos, these can be genre and metadata, and for music compositions, low-level audio features such as tonality, tempo, and timbre. This approach allows for effective recommendations of new releases and niche content, but creates the risk of forming a so-called 'filter bubble,' excessively narrowing the diversity of results.

The natural development of these two approaches has been the emergence of hybrid approaches that integrate their strengths while levelling out the weaknesses inherent in each individually. By combining user interaction analysis with a deep understanding of content attributes, hybrid systems achieve a new level of accuracy and stability, effectively coping with data sparsity and the cold start problem.

These theoretical approaches are implemented in practice using a wide range of machine learning algorithms. Classical methods, such as k-nearest neighbours (KNN)

[13], serve as the basis for calculating similarities in collaborative filtering approaches. More sophisticated techniques, including matrix factorisation algorithms [10] (SVD, ALS), allow the identification of hidden or latent factors behind user preferences, reducing a high-dimensional problem to manageable dimensions. The current stage of development is associated with the active introduction of deep learning methods. Neural Collaborative Filtering architectures model nonlinear interactions between users and content, while recurrent [1] (RNN) and convolutional [9] (CNN) neural networks open up opportunities for working with sequences of actions and complex multimodal features such as descriptions and audio waves.

The final and critically important element in the life cycle of a recommendation system is the evaluation stage [11]. The quality of prediction is traditionally measured by metrics of precision and recall, which are balanced by the F-measure. However, for real-world systems where recommendations are presented as a ranked list, these metrics are insufficient. Therefore, metrics that take into account the order of items, such as MAP (mean accuracy) and, in particular, NDCG (normalised discounted cumulative gain), come to the fore. NDCG reflects key business logic by assigning greater weight to relevant items at the top of the list, which directly correlates with user engagement and the overall success of the platform. Thus, the direction of further development and optimization of the system directly depends on the selection and interpretation of these metrics.

3 The Architecture of Recommendation Systems Using RuTube as an Example

RuTube is a Russian platform for watching and uploading videos, similar to YouTube, which focuses on personalising content in order to increase user engagement and retain its audience. The main functionality of the system is to provide recommendations on the home page, in playlists and in the 'next video' block, forming an individual feed for each user based on their previous interactions and preferences. The architecture of the recommendation system on RuTube is built using several approaches, each of which is focused on specific types of data and methods of processing them.

Collaborative filtering is one of the main methods of generating recommendations on the platform. It analyses user behaviour, including views, likes, subscriptions, and interaction history with videos, which allows it to identify similar patterns between users or between content objects. Within the item-based [14] approach, the system is able to recommend videos to users that are similar to those they have already watched, based on what videos other users with similar preferences like. This method is particularly effective for identifying hidden connections between popular content items and users, but may encounter limitations when new content or new users appear, requiring additional solutions to overcome the 'cold start' problem.

Content filtering complements the collaborative approach and relies on the characteristics of the objects themselves. For video content, these can be genres, key tags, descriptions and titles, as well as visual features extracted using convolutional neural networks [9]. The use of such features allows the recommendation system to generate suggestions based on the similarity of the video's theme and content, which is especially

important when working with new or unpopular content for which there is not yet sufficient user interaction statistics. Content filtering provides a more flexible approach to personalisation and helps to expand the variety of recommendations, while improving the quality of content selection for different audience categories.

To improve the accuracy of recommendations and minimise the limitations of each individual method, hybrid models are used that combine signals from collaborative and content filtering. In particular, the RuTube platform uses approaches based on the MatrixNet algorithm [7] developed by Yandex, which allows integrating various data sources and taking into account both behavioural and content features when ranking videos. This combination of methods allows the system to more accurately predict user interests, adjusting recommendations in line with current preferences and general trends in the user base. The process of data processing and recommendation generation on the platform involves several sequential stages. First, data about user interactions with the platform, including views, likes, viewing time, and other types of activity, is collected and stored. The data is organised and stored in distributed systems such as Hadoop [6] or Cassandra [2], which ensures scalability and the ability to work with large amounts of information. In the next stage, the data is converted into a user-item matrix, where rows correspond to users, columns correspond to video content, and values reflect the level of interaction, such as views, likes, or comments. This matrix serves as the basis for training machine learning models that can predict user interests and generate personalised recommendations. Model training is primarily conducted offline on large datasets using distributed computing environments such as Spark [5], which allows for the efficient processing of millions of interactions and the identification of hidden patterns. Once trained, models are applied online, providing real-time recommendations. The use of deep neural networks in video ranking allows not only user activity history to be taken into account, but also content characteristics such as visual features or semantics, ensuring high relevance. Thanks to the adaptability of the algorithms, the system quickly responds to changes in user preferences and the appearance of new videos, maintaining a stable level of engagement without compromising the quality of personalisation.

Thus, the architecture of the RuTube recommendation system is based on a hybrid approach that combines collaborative filtering and content-oriented methods. Implementation based on machine learning algorithms and distributed data processing technologies allows us to predict audience interests and maintain the scalability of the platform when working with large data sets.

4 The Architecture of Recommendation Systems Using Yandex. Music as an Example

Yandex.Music is an online platform for listening to and personalising music content, which allows users to find songs, playlists and radio stations according to their individual preferences. The main task of the recommendation system on the platform is to create personalised playlists, select songs based on previous listening history, and suggest new artists and genres that may interest the user. The system's architecture is built using a combination of collaborative, content-based, and hybrid approaches, which ensures high recommendation accuracy and adaptability to changing audience tastes.

Collaborative filtering on the Yandex. Music platform is based on the analysis of user interactions, including listening, likes, additions to playlists, and skips. Two approaches are used to recommend music: collaborative filtering based on users and filtering based on content. The first approach groups people with similar tastes. If other members of your 'music group' liked a certain song, the User-based filtering method [14] will suggest it to you as well. The second approach works differently: it analyses the tracks themselves. The algorithm finds patterns, for example, if users often listen to the same compositions together, it suggests similar music based on this [14]. The Item-based method [14] is good at recommending popular tracks, but it has a weakness: it does not work well with new music and new users. This problem is known as a 'cold start,' and additional methods are needed to solve it.

To make recommendations more accurate and solve the cold start problem, collaborative filtering is supplemented with content filtering. Here, the system no longer studies human behaviour, but rather the sound characteristics of audio recordings, such as genre, tempo, or tonality. For example, if a user frequently listens to energetic tracks of a certain genre, the system may suggest other compositions with similar characteristics, even if none of the user's friends or communities have listened to them before. To extract complex features, audio signal processing and machine learning methods are used, including convolutional neural networks (CNN) [9], which provide a more accurate understanding of track content and improve the quality of personalisation.

Hybrid models on Yandex.Music combine collaborative and content filtering signals and also use real-time user behaviour data. These models use recurrent neural networks (RNN) [1], which are capable of taking into account listening sequences and generating recommendations based on current preferences and context.

The formation of recommendations begins with data collection – the system records all user actions: which music tracks they listen to, add to their playlists or, conversely, skip. These data arrays (logs) are stored in distributed storage, which allows you to work with huge amounts of information. In the next stage, the collected data is structured into a matrix of 'user-object' interactions. In this matrix, the rows are users, the columns are music tracks, and the values in the cells are the level of interaction between them. For example, the latter can be expressed in the number of plays. This matrix becomes the main source of data for training machine learning models. It is used to build filtering algorithms and hybrid neural networks, whose task is to predict missing values in the matrix, i.e. to predict whether a user will like a particular composition that is unknown to them.

The models are trained offline on large data sets and then applied online to provide real-time recommendations. Recurrent neural networks allow the sequence of user actions to be taken into account, which is especially important for creating playlists and suggesting new tracks that match the user's current mood and preferences. In addition, hybrid approaches effectively address the 'cold start' problem by suggesting new tracks to users based on content characteristics and general patterns in the audience's musical preferences.

Thus, the architecture of the Yandex.Music recommendation system is a complex combination of collaborative, content-based, and hybrid methods implemented using modern machine learning algorithms, including deep neural networks. This approach

ensures high accuracy and personalisation of recommendations, allows adaptation to dynamic changes in user preferences, and enables efficient processing of large data sets, supporting the diversity and relevance of suggestions on the platform.

5 Comparative Analysis of RuTube and Yandex.Music Recommendation Systems

The RuTube and Yandex.Music recommendation systems are examples of platforms with different types of content — video and music, respectively — but with similar goals: increasing user engagement and personalising content. Despite the common concept, the architectures and algorithmic approaches of both systems have features determined by the specifics of the content being processed and the nature of user interactions.

RuTube is based on a combination of collaborative and content filtering [12] using hybrid models. The main focus is on analysing user views, likes and subscriptions to identify similar behaviour patterns and create personalised video feeds. Content filtering on the platform takes into account genres, tags, descriptions and visual characteristics of videos, which allows new items to be recommended even with limited interaction history. Model training is conducted offline using distributed data processing systems, and real-time recommendation generation is provided by deep neural networks capable of taking into account complex relationships between content and users. Yandex.Music, in turn, uses a similar hybrid approach, but with features specific to music content. Here, collaborative filtering analyses listening sequences, likes, skips and additions to playlists, while content filtering uses audio features of tracks, such as genre, tempo and tonality, extracted using machine learning and convolutional neural networks. An important feature is the use of recurrent neural networks, which allow listening sequences to be taken into account and recommendations to be dynamically adapted to the user's current preferences, which is especially important for creating playlists and suggesting new compositions.

The main differences between the platforms are related to the nature of the content and the specifics of user interactions. RuTube's video content is more varied in length, genre and format, which requires algorithms to take into account visual characteristics and video diversity. Yandex.Music's music content, on the other hand, is characterised by listening sequences and the significance of temporal patterns, which makes recurrent neural networks and audio feature analysis more effective for predicting user preferences. Both platforms use hybrid approaches to improve recommendation accuracy and overcome the 'cold start' problem, but the methods of integrating collaborative and content signals vary depending on the type of data and user activity characteristics.

In terms of scalability and processing large data sets, both platforms use distributed storage and processing systems. RuTube uses Hadoop [6] and Cassandra [2] to handle a huge number of video interactions, while Yandex.Music uses similar solutions for audio data and listening sequences. Recommendation quality metrics, including Precision, Recall, F1, MAP, and NDCG, are used on both platforms to monitor the accuracy and relevance of the content delivered, with the specifics of the data influencing the interpretation of results and the tuning of algorithms.

A comparative analysis shows that although RuTube and Yandex.Music use similar concepts of collaborative and content filtering, the specifics of content and user interactions determine the choice of specific algorithms, architectural solutions, and data processing methods. RuTube focuses on visual features and video diversity, while Yandex.Music focuses on audio features and listening sequences. In both cases, hybrid models and modern machine learning algorithms ensure high accuracy, personalisation, and adaptability of the systems, allowing them to maintain the relevance and accuracy of recommendations for users.

6 Problems and Prospects for the Development of Recommendation Systems

Modern recommendation systems, such as RuTube and Yandex.Music, demonstrate high efficiency in personalising content and increasing user engagement, but their operation is fraught with a number of difficulties and challenges that require constant improvement of algorithms and architectural solutions. One of the key problems is the 'cold start' problem that arises when a new user or new content object appears. In such cases, the lack of information about the preferences or characteristics of the object makes it difficult to form accurate recommendations. To solve this problem, hybrid approaches are used that employ both collaborative and content-based features, as well as clustering and preference prediction methods based on limited data.

Another significant problem is ensuring diversity and preventing 'filter bubbles' [8], where the user receives recommendations for a limited range of content, which can reduce interest and limit the discovery of new material. Solving this problem requires the implementation of diversity, ranking, and random sampling strategies that simultaneously take into account user preferences and expand the range of recommendations.

From a data processing perspective, modern platforms face challenges of scalability and high computational load, as they need to process millions of user interactions with large amounts of content in real time. To do this, distributed data storage and processing systems are used, as well as optimised learning algorithms, including offline and online training of machine learning models. Deep neural networks and hybrid architectures require significant computing resources, so it is important to strike a balance between the accuracy of recommendations and the efficient use of resources [3].

The prospects for the development of recommendation systems are linked to increased personalisation and the introduction of modern deep learning algorithms. The development of recurrent and transformer models makes it possible to take into account the sequence of user actions and more accurately predict their preferences, while generative approaches open up opportunities for creating new content and dynamic recommendations. Methods of model interpretability are also being actively researched, which allow explaining to the user why certain recommendations are offered, which increases trust in the system and the quality of the user experience. Additionally, with the growth of multi-platform services, recommendation systems will be integrated into more complex ecosystems where user behaviour data is collected from various sources, including social networks, streaming services, and mobile applications. This opens up new opportunities

to improve the accuracy of recommendations and create more personalised content, but at the same time requires solutions to data protection and user privacy issues.

Despite the existing challenges, recommendation systems continue to evolve, incorporating hybrid models, neural network algorithms, and scalable solutions for big data processing. These technologies not only improve the accuracy and diversity of recommendations, but also ensure long-term user engagement, setting the modern standard for content personalisation in online services.

7 An Illustrative Example of Collaborative Filtering in Python

To demonstrate the principle of user-based collaborative filtering, an example of implementation in Python using the scikit-learn library is provided.

The model generates a prediction of user interest in an object based on the similarity of user profiles, determined using the cosine similarity measure.

A small matrix of user-object interactions is used as the source data, where a value of 1 corresponds to an interaction (view, like, etc.), and 0 corresponds to no interaction (Fig. 1).

```
# Example of the simplest collaborative filtering using a user-item matrix
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity

# Matrix of user interactions with objects (1 – view/like, 0 – none)
# Rows – users, columns – videos or tracks
user_item_matrix = np.array([
    [1, 0, 1, 0],
    [0, 1, 1, 0],
    [1, 1, 0, 1],
    [0, 0, 1, 1]
])

# Calculating cosine similarity between users
user_similarity = cosine_similarity(user_item_matrix)

# Function for predicting a rating for a specific user and object
def predict_score(user_index, item_index):
    # We average ratings for similar users
    similar_users = user_similarity[user_index]
    item_ratings = user_item_matrix[:, item_index]
    # Exclude the user themselves
    weighted_sum = np.dot(similar_users, item_ratings) - similar_users[user_index]*item_ratings[item_index]
    similarity_sum = np.sum(similar_users) - similar_users[user_index]
    return weighted_sum / similarity_sum if similarity_sum != 0 else 0

# Example: predicting the first user's interest in the second object
predicted_score = predict_score(0, 1)
print(f"Predicted user interest 0 in object 1: {predicted_score:.2f}")
```

```
Predicted user interest 0 in object 1: -0.07
```

Fig. 1. The simplest collaborative filtering.

The algorithm calculates the cosine similarity between users and, based on this, assesses their interest in the selected object. The resulting value of −0.07 is close to zero, which indicates a low probability of user interest in the object, since there is no interaction with it among similar users. The example demonstrates the basic principle of user-based collaborative filtering: generating recommendations based on similarities

in user behaviour. Even a simple model with limited data allows us to predict interest in content without analysing its content characteristics. In real systems, such as RuTube and Yandex.Music, similar approaches are scaled using matrix factorisation, deep neural networks, and distributed computing systems, which significantly improves the accuracy and quality of personalisation.

8 Conclusion

This article examines the principles of recommendation systems using the example of Russian platforms RuTube and Yandex.Music, and demonstrates the application of machine learning algorithms for content personalisation. The theoretical foundations of collaborative, content-based, and hybrid filtering were analysed, and system architectures adapted to the specifics of video and music content were presented.

A comparative analysis demonstrated that the characteristics of the content and the nature of user interactions determine the choice of algorithms and architectural solutions. For RuTube, the use of visual features and video diversity is key, while Yandex.Music focuses on audio features and listening sequences. In both cases, hybrid approaches and neural network models make it possible to improve the accuracy of recommendations, take into account individual user preferences, and generate diverse personalised content.

The problems discussed — cold start, limited variety of recommendations, and high computational load — highlight the need for continuous improvement of methods and algorithms. Prospects for development include the introduction of transformer and recurrent models, generative approaches, improving the interpretability of recommendations, and integrating data from different sources. The illustrative example in Python showed how the basic principle of collaborative filtering can be implemented in practice, providing a clear understanding of the mechanism behind personalised recommendations. Despite the simplicity of the example, it reflects the fundamental idea behind modern recommendation systems, which rely on analysing user interactions and predicting their interests using machine learning methods.

Thus, the use of machine learning algorithms in recommendation systems ensures high accuracy, personalisation and adaptability of platforms, which is a key factor in increasing user engagement and the quality of the digital experience. Modern technologies open up vast opportunities for the further development and improvement of recommendation systems, making them more intelligent, flexible, and tailored to the needs of each user.

References

1. Alyoshin, N.A.: Recurrent neural networks. World Sci. Probl. Innov., 10–12 (2021)
2. Bohora, K., et al.: Backup and recovery mechanisms of Cassandra database: a review. J. Digit. Forensics, Secur. Law. **15**(2), 5 (2021)
3. Fomicheva, S.G.: Deep neural networks as an object of protection. Wave Electron. Info Commun. Syst., 223–228 (2021)
4. Knyazeva A.A. et al. Collaborative filtering for building recommendations based on order data//Bulletin of Novosibirsk State University. Series: information Technology 2018 16 2 62–69.

5. Kosenkov, V.V., Bogdanov, A.V.: Compromise between security and performance of Kubernetes container execution environments. In: Challenges of Modernity and Strategies for the Development of Society in the New Reality, pp. 145–153 (2022)
6. Ma, C., Zhao, M., Zhao, Y.: An overview of Hadoop applications in transportation big data. J. Traffic Transp. Eng. (Engl. Ed.). **10**(5), 900–917 (2023)
7. Mironov A., Khuziev I. Optimisation of decision tree evaluation using SIMD instructions. https://arxiv.org/abs/2205.07307. 2022.
8. Moro R., Pecher B. Auditing Youtube's recommendation algorithm for misinformation filter bubbles. 2022
9. Romanov, A.A.: Convolutional neural networks. Sci. Res. Key Probl. Third Millenn., 5–9 (2018)
10. Samokhina, I.V.: Research on matrix factorisation methods in recommendation systems. Int. Sci. Tech. Conf. Young Sci., 4015–4022 (2020)
11. Simanenko Yu.A. Evaluation of the accuracy of recommendation systems: research, problems and prospects BBK 1 N 34 2871
12. Sotnikova, E.A.: Recommendation Systems in Applications. Innovations Science. Educ. **35**, 511–519 (2021)
13. Tyart, N.A., Stepanov, N.N.: The k-nearest neighbours method and its software implementation for automated object classification. Mod. Technol. Sci. Education STNO, 7–10 (2016)
14. Yamashita, A., Kawamura, H., Suzuki, K.: Adaptive fusion method for user-based and item-based collaborative filtering. Adv. Complex Syst. **14**(02), 133–149 (2011)

Comparative Analysis of Artificial Intelligence Model Extraction Methods

Tatiana Peleshenko[1], Maxim Martemyanov[1(✉)], Zahid Raza[2], Ivan Borodulin[1], and Alexander Gish[1]

[1] North Caucasus Federal University, Stavropol, Russia
mixi26rus@mail.ru
[2] Jawaharlal Nehru University, New Delhi, India

Abstract. The paper presents a comparative analysis of artificial intelligence (AI) model extraction methods, focusing on model stealing attacks in full-access (white-box), limited-access (black-box), and partial-access (gray-box) scenarios. The authors examine key types of AI and machine learning (ML) models, including neural networks, deep learning, and generative models (GAN, VAE), along with their vulnerabilities. Major attack types are discussed: model extraction, data poisoning, backdoor attacks, and reverse engineering. Specific extraction methods are analyzed in detail, such as Accuracy Extraction Attack (white-box), Data Augmentation with Jacobian Matrix (white-box), Fidelity Extraction Attack (black-box), Adaptive Sample Selection (black-box), Algebraic and Hybrid Attacks (gray-box), and Hybrid Approach (gray-box). A comparative characterization is provided based on advantages, disadvantages, implementation, and applications. In the practical section, a Python program using TensorFlow and Keras is implemented for model extraction via the Fidelity Extraction method with Adaptive Sample Selection, applied to a convolutional neural network on the MNIST dataset. Results demonstrate high accuracy of the substitute model (~93.4% vs. ~94.4% for the original). Conclusions emphasize the need for enhanced protection of AI models against such attacks, highlighting the balance between efficiency and accessibility in gray-box methods.

Keywords: Model extraction · Machine learning · White-box attacks · Black-box attacks · Gray-box attacks · Accuracy extraction attack · Data augmentation with Jacobian matrix · Fidelity extraction attack · Adaptive sample selection · Algebraic and hybrid attacks for natural language models · Hybrid approach

1 Artificial Intelligence and Machine Learning Models as Objects for Data Mining

1.1 Definition and Main Types of AI and ML

Before moving on to methods of attacking artificial intelligence models, it is necessary to understand what we are dealing with. Artificial intelligence is a very broad concept in itself, most often meaning systems capable of making decisions based on big data

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 494–508, 2027.
https://doi.org/10.1007/978-3-032-27834-0_46

and complex algorithms [1]. And the most important part of such systems are machine learning models [2].

Artificial intelligence is a set of technological solutions that allows simulating human cognitive functions (including self-learning, searching for solutions without a predefined algorithm and achieving insight) and obtaining results comparable to at least the results of human intellectual activity when performing specific practically significant data processing tasks [3].

Machine learning (ML) is the process of optimizing model parameters using computational methods so that the behavior of the model reflects data and/or experience [3].

There are several types of AI models, starting with neural networks, which are based on the structure of the human brain. A neural network; an artificial neural network (neural network, NN, neural net, artificial neural network) is a network of two or more layers consisting of neurons connected by weighted connections with adjustable weights; each neuron receives input data and produces a result [3]. Modern neural networks can be trained on huge amounts of data and identify complex patterns in such a way that it is difficult for a person to even explain their work, but attackers do not need explanations – it is only important for them to use the vulnerabilities of these models for their own purposes.

Next, we will consider deep learning, multi-layered neural networks that are capable of not just learning from data, but independently creating multi-level hierarchies of knowledge. Deep learning (deep neural network learning) is an approach to creating extensive hierarchical representations by training neural networks with a large number of hidden layers. An approach to creating extensive hierarchical representations by training neural networks with a large number of hidden layers [3]. They are used almost everywhere: from the recognition of faces and voices to the diagnosis of diseases in medicine. On the one hand, these are huge opportunities, and on the other, these models are extremely sensitive to attacks, data manipulation, and unauthorized extraction.

Finally, generative models such as generative–adversarial networks (GAN) or variational autoencoders (VAE) deserve special attention. They are able to create new data that did not exist before – images, texts, or even audio tracks. If malicious actors extract such a model, they can not only gain access to confidential data, but also create convincing fakes and manipulate information on a completely new level, which can be used to harm a person.

The scientific literature discusses the role of these generative models and the risks associated with them. In particular, the authors Hailong Hu and Jun Pang of the scientific article «Stealing Machine Learning Models: Attacks and Countermeasures for Generative Adversarial Networks» clearly demonstrate how generative models can be extracted and then used to implement various attacks, including image and data manipulation [1]. Also in the article "Investigation of attacks on neural network models using data poisoning and the choice of methods of protection against them", the authors Petrov, A.V. And Payusova, T. I. The vulnerabilities of neural networks and the features of deep learning are analyzed in detail, which shows the need for stricter protection of such models [4].

Thus, understanding these basic types of AI models – neural networks, deep learning, and generative models – is necessary in order to understand the vulnerability of modern

intelligent systems, which will allow us to study threats and methods of attack–protection against the extraction of AI models.

Going deeper into AI and MO, there is a variety of approaches and models: from simple algorithms to the most complex multi-level neural networks. Let's start with simple models, which include such traditional approaches as linear regression, logistic regression, and decision trees: a simple formula that solves an understandable task: classify, predict, or predict an event. Simple models are often transparent and understandable, and therefore easier to become targets of attacks in order to extract logic.

Then there are more advanced approaches, such as the support vector machine (SVM) and ensemble methods (Random Forest, Gradient Boosting). These methods use more complex logic, they process data differently and are more resistant to simple manipulation. But in some cases, they are more dangerous, because successful extraction of such models allows attackers to manage much more serious financial transactions or medical diagnoses.

And finally, the most complex deep neural networks, including transformers and recurrent neural networks (RNNs). These systems can analyze huge data streams, identify deeply hidden patterns, and generate realistic images, texts, and audio. Today they drive cars, recommend music and movies, analyze financial markets and predict the development of diseases. It is more difficult to crack the model here, but if the attackers manage to extract them and use them for their own purposes, the consequences can be disastrous. Special attention should be paid to transformers. These models are currently used in the most advanced text generation and natural language analysis systems. Their uniqueness lies in the fact that they are able to take into account not just individual data elements, but also the context as a whole. Models based on the architecture of transformers are increasingly becoming targets for attacks aimed at extracting internal logic and manipulating the results of their work [5].

That is why it is so important to take into account the full variety of machine learning models, from the simplest and most transparent to the most complex and multidimensional. Only a complete understanding of the specifics and vulnerabilities of each group of models will allow us to create a truly reliable protection system and prevent the most diverse and dangerous attacks on intelligent systems.

1.2 Main Types of Attacks on AI Models

Let's look at the main types of attacks on AI models:

(1) Extraction attacks are threats that make all efforts to train and protect machine learning models vulnerable. Unlike conventional attacks on data, which can be aimed at data manipulation or theft, attacks on model extraction allow attackers to restore not only the structure of the model, but also its internal architecture, logic, parameters, as well as its decision-making methods. An attacker can not only find out how the model works, but also use it for their own purposes.

The basic principle of model extraction attacks is that the attacker interacts with the model through its public interface – API or other available mechanisms, sending requests and analyzing the model's responses. Using multiple queries, an attacker can gradually reconstruct the structure of the model and its behavior, and then reproduce the

model by reusing or modifying it. One of the most well-known examples of such attacks is model extraction attacks, when malicious intruders use a query system to reconstruct models trained on large datasets. Attacks can be directed at both generative models and classifiers [6].

As a result of model extraction attacks, attackers can not only gain access to the model itself, but also use it to generate new data, which poses a serious threat to systems such as content generation, medicine, financing, etc., which undermines confidence in AI models.

(2) Attacks on data (data poisoning, backdoor attacks) are another class of threats that affect not only the extraction of models, but also their training. In particular, data poisoning attacks and backdoor attacks can cause the model to malfunction or intentionally produce malicious results.

Data poisoning consists in the fact that an attacker intentionally introduces errors or false data into the training set of a model, which distorts its behavior. For example, this can happen in the field of medicine, when a model is trained on distorted patient data, which leads to incorrect diagnoses. Backdoor attacks, unlike data poisoning, are aimed at introducing hidden vulnerabilities into the model that are activated only under certain conditions. This can be a special trigger that, when entering certain data (for example, images with a certain pattern), causes the model to produce pre-programmed errors or manipulate fake results [7].

(3) Attacks on the architecture of the model, or reverse engineering, are aimed at extracting not only model data, but also its internal structural elements, such as parameters of neural network layers or algorithms for their interaction. Unlike previous types of attacks, where attackers can only change the behavior of the model, with reverse engineering they gain access to its full-fledged internal structure, which allows not only duplicating the model, but also changing it for further attacks.

Attacks of this type are especially dangerous in the case of models based on complex neural networks or deep generative models, where the slightest change in architecture can greatly change the behavior of the entire system. In the case of GAN or VAE, for example, such attacks allow not only to extract information about the parameters, but also to modify the model to create fake images or generate false data. Thus, attacks on model logic extraction and reverse engineering pose information security threats that violate intellectual property and data security.

1.3 Overview of Black Box and White Box and Gray Box Attacks

Let's consider approaches to attacks aimed at extracting models:

1. Attacks on a model with full data access (white–box) are attacks in which an attacker has full access to all internal data of the model, including its parameters, architecture, and training set. This situation creates the greatest risks, because an attacker can not only restore the model, but also fully understand its logic, which makes it vulnerable to attacks and manipulations.

One of the most well-known methods of the white box attack is the use of gradient descent and gradient methods to reconstruct the model. For example, the Fast Gradient Sign Method (FGSM) is a method that uses information about the gradients of a model to create adversarial examples. Such examples lead the model astray, forcing it to draw erroneous conclusions. This method is widely used in the field of machine learning, as it allows you to effectively create examples that can be used to attack the model. The FGSM method allows an attacker to calculate the gradients of the model relative to the input data and then make minor changes to this data in order to cause an erroneous classification. This is done by optimizing the loss function using the gradient sign, which creates an insignificant but very effective perturbation for the model. Thus, an attack can be carried out even with limited access to data, if the attacker has access to the architecture of the model and its training [8].

2. Attacks on the model without access to internal data (black–box) are characterized by the fact that the attacker does not have direct access to the internal data of the model, such as its parameters, weights, or algorithm. In the case of black box attacks, the attacker uses the external interface of the model, usually an API or public queries, to restore the model or manipulate its behavior.

One of the most common methods of attacks in the black box is Genetic Algorithms (GA). In these attacks, the attacker tries to create an optimized model by generating and evaluating a population of possible models, choosing the most effective ones and repeating this process. Genetic algorithms can be especially effective for attacking complex models when simple query extraction is not possible. Another approach to black box attacks is Zero–Query Attacks, which have become popular in recent years. The attacker uses these attacks to manipulate the results of the model without having to edit the queries. Instead, the attacker tries to guess (predict) the parameters of the model based on external observations of its behavior, which makes such attacks even more dangerous and difficult to defend against [9].

3. Methods for attacks on models with partial data access of the gray box type focus on limited access to the model and require the creation of copies of models or the use of intermediate data. These methods are suitable for real-world conditions where an attacker cannot gain full access to the model, but can still use similar data or models. However, their effectiveness strongly depends on the quality of the available data and the accuracy of the model copies. Gray box attacks assume that the attacker has partial access to the model. This may include access to its output data, intermediate layers, or parts of its architecture. In this case, the attacker can use a combination of white and black box techniques, which makes these attacks more powerful than black box attacks, but less powerful than white box attacks. Gray box methods combine the analysis of the model's output data with some internal knowledge about its structure, which allows attackers to extract the model with lower computational costs [10].

For example, a Semi–Black–Box Attack is an attack in which the attacker has partial access to the target model, for example, access to its outputs or intermediate output points, but not to its internal parameters or training data. This type of attack is well suited for situations where the model is trained in a cloud environment or used in a system with limited access. The attacker has access to the model's output or intermediate

results, and uses this information to create adversarial examples that can influence the model's decision. Unlike white–box attacks, in a Semi–Black-Box attack, the attacker can use the results of the model's conclusions to adapt his attack, which allows him to find vulnerabilities in the model. Or a Semi–supervised Learning attack, in which the attacker uses partially labeled data to extract the model, which allows him to train the model with a limited amount of labeled data, which reduces training costs [11].

Conclusions the analysis revealed that black-box attacks offer the greatest versatility and stealth but require a significant number of requests, while white-box attacks offer high accuracy but are limited by the need for full access to the mod-li, making them preferable for insider threats.

2 Review and Comparative Characterization of AI Model Extraction Methods

2.1 Method 1: Accuracy Extraction Attack, White Box

The attacker has full access to the target model and uses its internal data to accurately restore its functionality and parameters, creating an identical replacement model [12].

The extraction algorithm proceeds as follows. Gain access to the architecture, parameters, and training data of the target model, for example, through insider access or leakage. Next, use this data to train an identical replacement model by copying weights and structure. After that, you need to check the accuracy of the replacement model on the test data, comparing the predictions with the target model.

The Accuracy Extraction Attack method has several advantages. The replacement model can accurately reproduce the behavior of the target model, achieving almost 100% accuracy, it is also possible to obtain accurate model parameters, which allows you to completely duplicate its functionality and full access eliminates the need for a large number of queries, which speeds up the process.

However, there are disadvantages to this method. It requires full access, which is rarely possible for proprietary models, as it requires insider access or data leakage, and it also requires a deep understanding of the internal structure of the model, which can be technically difficult to implement. The main disadvantage of the method is the high risk of detection: if access is the result of a leak or unauthorized access, this can lead to catastrophic consequences.

The method is applicable in research environments or in the presence of insider access, for example, to improve open models or create competing products in academic settings.

2.2 Method 2. Data Augmentation with Jacobian Matrix, White Box

The method uses the Jacobi matrix of the target model to generate adversarial examples, which are then used to train the replacement model, ensuring high accuracy of reconstruction [13].

The AI model is extracted using some algorithm. First, you need to calculate the Jacobi matrices of the target model relative to the input data using gradients relative

to the inputs. After that, you need to generate adversarial examples close to the decision boundary using these gradients. And finally, train the replacement model on these examples and source data, minimizing the discrepancy with the target model.

The advantages of the Data Augmentation with Jacobian Matrix method are: high portability: adversarial examples are effective for different models, which makes the method universal. There is also increased accuracy, which helps to create a more accurate replacement model, especially for neural networks.

One of the disadvantages of this method is that calculating the Jacobi matrix requires significant resources, which may be a limitation for large models. The method also requires full access: it is necessary to know the gradients of the model, which limits the application.

The method is useful for creating stable replacement models or analyzing the vulnerabilities of the target model, for example, in security systems where full access is possible, such as internal corporate systems.

2.3 Method 3. Fidelity Extraction Attack, Black Box

In a black box environment, the attacker has access only to the inputs and outputs of the model (for example, via the API), but does not know its internal structure or parameters. The attacker sends multiple requests to the target model via the API, collects output data (for example, predictions or generated images for GAN) and uses them to train a replacement model that mimics the behavior of the target model. The method does not require knowledge of architecture or model parameters [1].

The algorithm for extracting a model with "black box" access is as follows. First, you need to send multiple requests to the target model, collecting input data (for example, images or text) and corresponding outputs (predictions or generated data). Secondly, we will use the collected data to train a replacement model, for example, a new GAN that reproduces the distribution of outputs of the target model. As a result, it is necessary to evaluate the quality of the replacement model using metrics such as Frechet Inception Distance (FID) for generative models or classification accuracy for discriminative models.

The advantages of the Fidelity Extraction Attack method, first of all, are the minimum requirements: you only need access to the API of the model, which makes the method accessible to a wide range of attackers. The method is also applicable to any models available through queries, regardless of their architecture. It is also more difficult to detect, as it uses standard queries, which reduces the risk of blocking.

The disadvantages of this method are that it requires a large number of queries (for example, 50,000 for GAN) to achieve acceptable accuracy, which may be a limitation in real models. Also, the replacement model may not perfectly reproduce the behavior of the target model, especially for complex models such as deep neural networks. In addition, only functionality is recreated, not the exact parameters, which limits the possibilities for further analysis.

The method is used to steal the functionality of models in cloud services such as face recognition (Face++) or image generation, where API access is limited. An example is the theft of an image generation model in services like DALL·E to create competing products.

2.4 Method 4. Adaptive Sample Selection, Black Box

This method uses unlabeled data from related domains and active learning to select the most informative samples, which are then requested from the target model. The responses are used to train a replacement model, which reduces the total number of requests [14].

The extraction algorithm is as follows. Unlabeled data is collected from the domain associated with the target model, for example, images from similar datasets if the model classifies images. Next, active learning is used to select the most informative samples for queries to the target model, using strategies such as selecting data with the greatest uncertainty. Finally, a replacement model is trained based on the model's responses, minimizing the discrepancy between the predictions of the target and replacement models.

Some advantages of this method can be noted: reducing the number of queries by selecting the most informative data, which makes the method economical in terms of resources. The method can also achieve high accuracy with fewer queries compared to simple queries, which improves the quality of the replacement model.

There are also some disadvantages of this approach. The method requires the use of active learning and knowledge of related domains, which increases the complexity of the implementation. Access to non-labeled data is also required, which may be a limitation in some models, especially if suitable data is not available.

The method is suitable for cases where data similar to the training data of the target model is available, for example, in image classification tasks using data from other sets, such as CIFAR-10 for a MNIST-trained model.

2.5 Method 5. Algebraic and Hybrid Attacks for Natural Language Models, Gray Box

In a gray box environment, the attacker has partial knowledge of the model, for example, its architecture or some parameters, but not full access. This method uses partial knowledge of the model, for example, a public pre-trained encoder (for example, BERT) and a private classifier. Algebraic methods restore the classifier parameters with a minimum number of queries, while hybrid approaches combine algebraic and learnable techniques [15].

Implementation algorithm: 1. Use a public encoder such as BERT to generate embeddings for a set of input data, for example, text sentences. 2. Request classifications from the target model for these embeddings using API or partial access. 3. Apply algebraic methods to solve the system of equations and reconstruct the classifier parameters using the answers obtained. 4. Combine with learnable methods, such as minimizing discrepancies, if necessary, to improve accuracy.

The method has a number of advantages: 1. Requires fewer queries than purely trainable methods, which reduces the load on the API. 2. Achieves high accuracy for models with public encoders, especially in natural language processing tasks. 3. Designed for large natural language processing models such as BERT, which makes it specialized.

However, there are several significant disadvantages of the Algebraic and Hybrid Attacks for Natural Language Models method.: 1. The method only works for models with a public encoder and a private classifier, which limits its use. 2. Requires partial

knowledge: it is necessary to know the architecture of the encoder, which may not be available in some cases.

The method is suitable for models where part of the architecture (for example, the encoder) is public, for example, in mood analysis tasks using BERT, especially in academic or corporate environments with partial access.

2.6 Method 6. Hybrid Approach, Gray Box

It combines black and white box techniques using partial knowledge of the model, for example, access to generated data, latent codes, or assumptions about the model structure, but not to the full architecture or discriminator [16].

Implementation algorithm: 1. Use available data such as generated samples or latent codes obtained from partial access. 2. Make assumptions about the model, for example, about the distribution of latent codes, based on a known part of the architecture. 3. Train the replacement model by combining API requests (as in a black box) and partial knowledge (as in a white box), minimizing the discrepancy with the target model.

The method has a number of advantages: 1. Adapts to the level of available knowledge, which makes the method universal. 2. Less noticeable than white box attacks, which reduces the risk of detection. 3. It may be more efficient than a black box, but simpler than a white one, offering a compromise between access and accuracy.

However, there are several significant disadvantages of the Hy-brid Approach method: 1. Effectiveness depends on the amount of information available, which can lead to unstable results. 2. The method requires a combination of various techniques, which increases the complexity of the implementation. 3. Limited accuracy: lower than that of the white box due to the lack of full access.

The method is suitable for models with partial access, for example, in academic settings where part of the model may be publicly available, or in corporate environments with limited access to internal data.

2.7 Comparative Characterization of AI Model Extraction Methods

Methods for extracting AI models in a white box environment are the most accurate and efficient, but are limited by the need for full access to the model, which makes them difficult to implement and vulnerable to detection. Implementation algorithm: 1. Use available data such as generated samples or latent codes obtained from partial access. 2. Make assumptions about the model, for example, about the distribution of latent codes, based on a known part of the architecture. 3. Train the replacement model by combining API requests (as in a black box) and partial knowledge (as in a white box), minimizing the discrepancy with the target model.

Black box methods are versatile and secretive, but they require a large number of queries and may not achieve full accuracy, which reduces their effectiveness for complex models. Gray box methods offer a balance between accuracy and accessibility, but depend on partial knowledge of the model, which makes them applicable in non-standard tasks with limited access. A comparative analysis of existing methods for extracting AI models is presented in Table 1. Based on a comparison of the previous paragraphs.

Table 1. Comparative characterization of AI model extraction methods.

Mailbox type	Method	Advantages	Disadvantages	Implementation method	Application
White box	Accuracy extraction	High accuracy, parameter recovery, fewer queries	Requires full access, complexity, risk of discovery	Copying parameters, training an identical model	Research environments, insider access
White box	Data augmentation with jacobian	High tolerance, accuracy	Computational complexity, full access	Generation of adversarial examples, model training	Vulnerability analysis, enterprise systems
Black box	Fidelity extraction	Minimum requirements, versatility, secrecy	Inefficiency, limited accuracy, no parameters	API requests, training of the replacement model	Cloud services, image generation
Black box	Adaptive sample selection	Efficiency, high accuracy	Complexity, requires data	Active learning, selection of in-formative inquiries	Classification with similar data
Gray box	Algebraic and hybrid attacks	Effectiveness, accuracy, applicability to NLP	Limited, partial knowledge	Encoder use, algebraic methods	Models with public encoders, sentiment analysis
Gray box	Hybrid approach	Flexibility, stealth, balance	Variable efficiency, complexity	Combination of queries and partial knowledge	Academic scenarios, partial access

Based on a comparative analysis of various AI pattern extraction attack methods, it was found that they pose a serious threat to security and intellectual property. Black-box methods such as Fidelity Extraction and Adaptive Sample Selection are versatile but require multiple queries, which can be a limitation. White-box methods, such as Accuracy Extraction and Data Augmentation with Jacobian Matrix, provide high accuracy but are limited by the need for full access, which is rarely possible. Gray boxes, including Algebraic Attacks and Hybrid Approach, offer balance but depend on partial knowledge, making them effective in specific scenarios. Practical implementation shows that even with limited access it is possible to recreate the functionality of the model, highlighting the need for further research in security.

Conclusions. The Comparative Characterization Reveals that Gray-Box Methods Offer the Best Balance Between Efficiency and Accessibility, while White-Box Methods Offer High Accuracy but Require Full Access. As a Result, Gray-Box Methods Are the most preferred for AI Model Extraction.

3 Implementation of a Program for AI Model Extraction

3.1 Rationale for the Choice of Method for the Study

The Fidelity Extraction method with Adaptive Sample Selection was chosen for implementation in this paper for the following reasons:

- The method works with limited access to the model, using only input data and predictions (API–like access), which corresponds to real-world scenarios where the internal structure of the model is not available.
- The method allows you to create a replacement model that accurately mimics the behavior of the target model, as shown in the results (the accuracy of the replacement model is ~93.40% versus ~94.42% of the original one).
- The method does not require complex mathematical transformations or access to the gradients of the model, which simplifies its implementation using standard libraries such as TensorFlow and Keras.
- The method is suitable for various types of deep learning models, including convolutional neural networks (CNNs) used in the image classification task (MNIST).
- Adaptive Sample Selection integration allows you to select the most informative samples based on the entropy of predictions, which increases the effectiveness of the attack with a limited number of requests.

3.2 Description of the Implementation Algorithm and Creation of the Program

To implement the extraction of the AI model, the Fidelity Extraction method with Adaptive Sample Selection was used, which allows you to create a replacement model that mimics the functionality of the target model without access to its internal structure. The method is based on collecting predictions of the target model for a set of input data and then training a replacement model based on these predictions. Adaptive sample selection increases efficiency by selecting the most informative data based on uncertain predictions.

To begin with, we will implement the convolutional neural network itself, which we will attack. The method is used to analyze the vulnerabilities of AI models, demonstrating the possibility of extracting their functionality in conditions of limited access, which is important for assessing the security of models used in the API. In this case, the program works with the MNIST dataset (dataset: MNIST; first activation function: ISReLU; second activation function: Softsign; convolutional layers: 2; hidden layers: (3) and user-generated images created through a graphical interface that mimics the API. The implementation algorithm is shown in Algorithm 1.

Algorithm 1: Example of creating a model with the ISRLU activation function

```
Input: x, alpha=1.0
Output: Returns the result of applying the ISRLU function
Function isrlu(x, alpha=1.0);
begin
  return if where(x ≥ 0, x, x / sqrt(1 + α · tf.square(x)));

Creating a model:

model = Sequential([
  Conv2D(filters=1, kernel_size=(3, 3), input_shape=(28, 28, 1), padding='same',
activation=isrlu);
  MaxPooling2D(pool_size=(2, 2));
  Conv2D(filters=1, kernel_size=(3, 3), padding='same', activation=isrlu);
  MaxPooling2D(pool_size=(2, 2));
  Flatten();
  Dense(32, activation='softsign');
  Dense(16, activation='softsign');
  Dense(16, activation='softsign');
  Dense(10, activation='softmax');
])
```

The model extraction algorithm (see Algorithm 2):

- The MNIST dataset is used, which contains images of handwritten digits (28x28 pixels) converted to the form (N, 28, 28, 1). Then the user images are downloaded from the file weights.txt, created through a graphical interface. If the file is missing or incorrect, MNIST test data is used.
- The target model is a convolutional neural network (CNN) with two convolutional layers using custom isrlu activation and fully connected layers. The model is trained on MNIST training data (60,000 images) for 10–15 epochs, achieving an accuracy of ~94–95% on the test data.
- 1000 images are selected for the attack (from weights.txt or the MNIST test suite). Predictions of the target model (y_pred) are obtained for these images by simulating API requests.
- The entropy of predictions is calculated for each image and 1000 images with the highest entropy (the most uncertain predictions) are selected to maximize the informative value of the data.
- A replacement model is CNN with 32 filters in a convolutional layer, one MaxPooling layer and fully connected layers. Next, the Model is trained on the selected data and their predictions from the target model for 10–15 epochs.
- The accuracy of the replacement model is evaluated on MNIST test data and compared with the accuracy of the target model.

Algorithm 2: The main algorithm for extracting the AI model

Input: File weights.txt (user images), test MNIST data (X_test, y_test), target model (model), number of queries (num_queries=1000)
Output: Accuracy of the substitute model and the original model

Main Algorithm:

```
x_unlabeled = load_images_from_weights("weights.txt"); // Load user data
if x_unlabeled.size > 0 then
    x_unlabeled = x_unlabeled[:num_queries];
else
    x_unlabeled = X_test[:num_queries]; // Use MNIST test data

y_pred = model.predict(x_unlabeled); // Get predictions from the target model
y_pred_cat = keras.utils.to_categorical(np.argmax(y_pred, axis=1), 10);
x_unlabeled, y_pred_cat = select_adaptive_samples(x_unlabeled, y_pred, num_queries); // Adaptive sample selection based on entropy

substitute_model = Sequential([
    Conv2D(32, kernel_size=(3,3), activation='relu', input_shape=(28,28,1)), // Convolutional layer
    MaxPooling2D(pool_size=(2,2)), // Max pooling
    Flatten(), // Flatten into a 1D vector
    Dense(128, activation='relu'), // Fully connected layer
    Dense(10, activation='softmax') // Output layer
]);

substitute_model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy']); // Compile the model
substitute_model.fit(x_unlabeled, y_pred_cat, epochs=10, batch_size=32); // Train the substitute model

test_loss_sub, test_accuracy_sub = substitute_model.evaluate(X_test[:num_queries], keras.utils.to_categorical(y_test[:num_queries], 10)); // Evaluate accuracy
print("Accuracy of extracted substitute model: {test_accuracy_sub*100:.2f}%");
print("Original model accuracy: {test_accuracy*100:.2f}%");
```

Algorithm of drawing with the help of interface is realized with the help of tkinter library. The results are presented in Fig. 1.

Adaptive Sample Selection is a technique in machine learning in which the process of selecting data for training or analysis is dynamically adjusted based on current results or criteria in order to increase the effectiveness, accuracy or sustainability of the model. Unlike random or fixed selection, the adaptive approach allows you to focus on the most informative samples, minimizing the amount of data used and speeding up learning.

For example, in label noise learning tasks, adaptive sample selection uses mini–batch statistics to select reliable data, which improves the robustness of the model to noise [1].

4 Conclusions

The Chapter Describes the Implementation of a Program for AI Model Extraction Using the Fidelity Extraction Method with Adaptive Sample Selection. It First Justifies the Choice of the Method, which Is Recognized to Be Effective for Black Box Due to

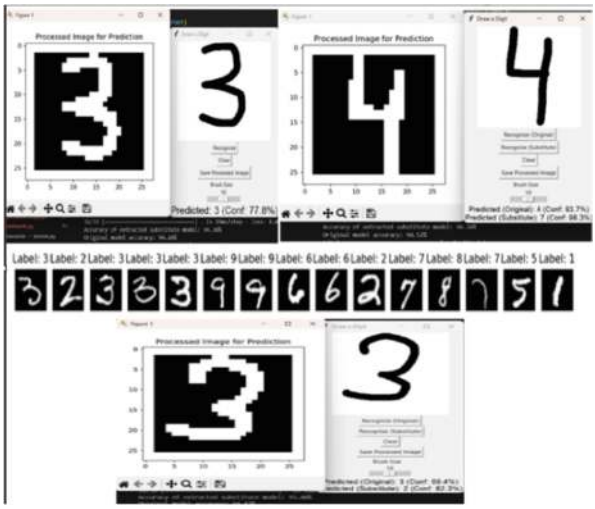


Fig. 1. Results.

its Simplicity, Versatility and High Accuracy of the Extraction. The Method Extracts the Model Functionality with an Accuracy of ~93.40%, which Is Close to the Original (~94.42%). The Algorithm of Model Extraction and the Process of Program Creation, Including Training of the Target Model, Selection of Informative Samples, Training of the Replacement Model and Development of a Graphical Interface for Interaction with the User Are Discussed in Detail below. The Program Is Implemented in Python Using the TensorFlow and Keras Libraries, which Ensures its Accessibility and Reproducibility. The Graphical Interface Allows Comparing the Predictions of the Original and Replacement Models, Demonstrating the Success of the Attack. The Implementation Supports the Use of Custom Images, which makes it Flexible and Relevant to Real-World Application Scenarios.

References

1. Hu, H., Pang, J.: Stealing machine learning models: attacks and countermeasures for generative adversarial networks. In: Proceedings of the 37th Annual Computer Security Applications Conference, pp. 1–16 (2021)
2. Chistyakov A., Andreev A. Attacks on Artificial Intelligence. https://media.kaspersky.com/ru/business-security/attacks-on-artificial-intelligence-whitepaper.pdf. Accessed 24.04. 2025. (2019)
3. GOST R. 71476–2024. Artificial Intelligence. Concepts and terminology of artificial intelligence, official site [Electronic resource]. https://gostassistentru/doc/d5025070–e226–4d6b–9871–ad5721156f8d. Accessed 15.04.2025
4. Petrov A.V.V., Payusova T.I. Investigation of attacks on neural network models with the help of "poisoning" data and the choice of methods of protection against them. Mathematical and Information Modeling: Proceedings of the All-Russian Conference of Young Scientists. Vyp. 21.-Tyumen, 2023. TyumSU-Press, (2023)

5. AI vulnerability: types of LLM-injection attacks and ways to protect against them. Hitech 2024. [Electronic resource]. https://hightech.fm/2024/11/21/llm-injection. Accessed 01.05.2025
6. Tramèr, F., et al.: Stealing machine learning models via prediction {APIs}. In: 25th USENIX security symposium (USENIX Security 16), pp. 601–618 (2016)
7. Artificial Intelligence Games: Attacks on Machine Learning Models and Their Consequences. CRIPTONIT 2022 [electronic resource]. https://hightech.fm/2024/11/21/llm-injection. Accessed 11.05.2025
8. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572 2014
9. Liang, J., et al.: Model extraction attacks revisited. In: Proceedings of the 19th ACM Asia Conference on Computer and Communications Security, pp. 1231–1245 (2024)
10. Kotenko, I.V., et al.: Attacks and defense methods in machine learning systems: analysis of modern research. Voprosy cybersecurity. (1), 59 (2024)
11. Semi-Supervised Learning in Machine Learning in simple words. dzen.ru 2021. [Electronic resource].https://dzen.ru/a/YDUxsqMy3XNzDoUh. Accessed 30. 05. 2025
12. Farid Ahamed S. et al. Evaluating query efficiency and accuracy of transfer learning–based model extraction attack in federated learning 2025. https://arxiv.org/abs/2505.23791v1.
13. Papernot, N., et al.: Practical black–box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519 (2017)
14. Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: stealing functionality of black–box models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4954–4963 (2019)
15. Santiago, Z–B., et al.: Grey–box extraction of natural language models. In: International Conference on Machine Learning. PMLR, pp. 12278–12286 (2021)
16. Béguelin, S.Z., et al.: Grey–box extraction of natural language models. ICML, (2021), 12278–12286
17. Patel, D., Sastry, P.S.: Adaptive sample selection for robust learning under label noise. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3932–3942 (2023)

The State of Quantum Machine Learning: Architectures, Encoding Strategies, and Generalization Challenges

Alicia Theresa Pereira, Sharlene Anna Pereira, Melsa Magdalene Fernandes, and Mary Anita E A(✉)

Christ University, Bengaluru, India
maryanita.ea@christuniversity.in

Abstract. Quantum Machine Learning (QML) is an emerging paradigm at the intersection of quantum computing and machine learning. It holds the potential for substantial computational accelerations and more powerful modeling. This survey provides a comparative review of recent advances in the subject, with emphasis placed upon algorithms, architectures, and applications. The discussed studies present developments in hybrid quantum-classical models, variational quantum circuits, and quantum neural networks, as well as applications to classical and quantum-native data. Key findings include the capacity of quantum algorithms to be more effective than classical methods in classification and clustering problems, the importance of data encoding and error mitigation, and the need to enhance trainability and generalization in the Noisy Intermediate-Scale Quantum (NISQ) regime. There are still challenges in relation to scalable dataset availability, preparing the state efficiently, and automated circuit design. This survey captures the current state of QML, highlighting both the theoretical potential and the necessity for practical approaches that can make its deployment in real-world applications scalable.

Keywords: Quantum machine learning · Variational quantum circuits · Quantum neural networks · NISQ · Quantum architecture search · Hybrid quantum-classical · Trainability · Generalization

1 Introduction

Quantum Machine Learning (QML) is a new field that integrates quantum computational concepts into machine learning methods. Leaning on quantum effects such as superposition and entanglement, QML seeks to surpass the computational complexity of traditional algorithms, particularly in the processing of high-dimensional data and challenging optimization problems. The rapid progress of quantum hardware, even in the noisy and limited NISQ era, renders QML a prime and extremely promising field of research in both machine learning and quantum computing.

Three dominant themes underlie current research in QML: architecture design, encoding methods, and generalization. Architecture design encompasses quantum neural networks, quantum convolutional models, and hybrid quantum-classical variational

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 509–516, 2027.
https://doi.org/10.1007/978-3-032-27834-0_47

circuits, which form the computational underpinning of QML systems. Encoding methods describe how data is encoded in quantum states from classical or quantum-natively optimized data, a process that significantly affects trainability, accuracy, and potential for quantum advantage. Generalization and trainability focus on how QML models learn from sparse data, avoid optimization failure phenomena such as barren plateaus, and extrapolate to novel tasks. This review provides a comparative overview of recent progress on these fronts, highlighting both theoretical potential and practical limitations while outlining the directions that are most significant for achieving scalable, real-world applications.

The lack of a proper organization in the field of information security poses a huge threat, as it can lead not only to financial losses, but also to reputational damage and legal consequences [11]. In this regard, information security issues are extremely relevant today.

2 Literature Review

Quantum Machine Learning (QML) has been investigated from various angles, namely theoretical development, algorithm design, and application on NISQ devices. Jadhav et al. [1] discuss the promise of QML for traditional data applications, wherein it sees exponential speed-ups in QSVM and Quantum k-means but points out challenges such as state preparation and hardware noise. Abohashima et al. [2] broadens this by categorically placing QML techniques as being entirely quantum, quantum-inspired, and hybrid paradigms, explaining encoding techniques, and emphasizing the contribution of error mitigation towards robust performance. Wossnig [3] takes into account classical data sets, discussing supervised QML performance in terms of data encoding, measurement overhead, and the feasibility of exponential speed-ups under low-rank or sparse regimes. Together, these papers give a glimpse into the promise of QML for classical data but also highlight pragmatic roadblocks.

The other research addresses quantum-native data and blended approaches. Schatzki et al. [4] present entangled quantum datasets and show that QNNs and QCNNs can effectively learn and classify multipartite entangled states while avoiding embedding challenges and barren plateau issues one faces when applying classical data. Baughman et al. [5] examine feature importance in QML models deployed on real-world datasets, coupling classical explainability techniques with restricted-qubit quantum models. Zaman et al. [6] give an extensive survey of hybrid quantum-classical approaches, such as variational quantum circuits, kernel methods, and point out research holes such as the requirement for standardized benchmarks, co-design techniques, and scalable hardware.

Other work targets bringing forward architectures and optimization during the NISQ era. Qi et al. [7] explore hybrid frameworks combining classical tensor networks and pretrained models with variational circuits through reinforcement learning and generative AI for quantum circuit architecture search (QCAS). Chen and Liang [8] offer a comprehensive tutorial on variational quantum circuits and quantum architecture search, offering applications such as QLSTM, federated QML, and quantum-assisted model compression. Caro et al. [9] offer rigorous generalization bounds for QML, demonstrating that small amounts of training data can be enough under efficient parameterizations,

with use cases in phase classification and unitary compiling. Thanasilp et al. [10] explore trainability problems, such as barren plateaus and dataset-induced barren plateaus, and how input state entanglement and encoding selection have a crucial impact on gradient behavior and model convergence (Fig. 1).

Hybrid quantum-classical methods predominate present research [2, 6–8], providing useful means to exploit NISQ devices. Variational quantum circuits (VQCs) and quantum neural networks (QNNs) are at the core of algorithm design [4, 7, 8]. Encoding and choice of dataset, either classical or quantum-native, is consistently cited as an important factor in determining trainability and generalization [3, 4, 9, 10]. Research gaps persist on a consistent basis on scalable dataset generation, optimal circuit construction, state preparation, and theory-practice bridging of speed-ups [1, 6–8]. Combined, these works frame a discipline advancing swiftly in theory and algorithm construction, yet still hampered by equipment restrictions, trainability issues, and the necessity for pragmatic verification (Table 1).

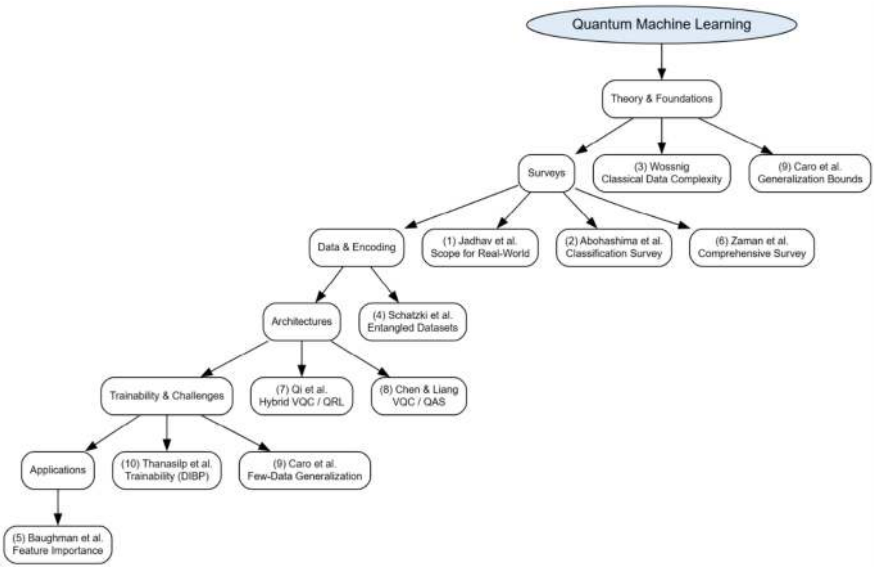


Fig. 1. Hierarchical tree structure categorizing quantum machine learning research.

3 Methods

The materials and methods used throughout the review work can be categorized into three broad types: datasets, quantum models and algorithms, and training/optimization strategies.

Datasets: Classical and quantum-native datasets are used. Classical datasets, e.g., those for text, image, or tabular classification [1–3], are quantized into quantum states using amplitude, angle, or basis encoding methods. Quantum-native datasets [4], e.g., the NTangled dataset, offer controlled multi-partite entangled states so that models can be trained directly from quantum states without embedding overhead. Hybrid research

Table 1. Comparison of studies.

Paper	Focus area	Technique	Key findings
Jadhav et al.	Real-world QML scope	Review of QSVM, k-means, QNNs	QML promises speed-ups; limited by state prep and NISQ noise
Abohashima et al.	QML classification	Survey: Pure quantum, quantum-inspired, and hybrid quantum-classical	Encoding central; phased scheme; noise/error key challenge
Wossnig	Classical data QML	Complexity analysis	Exponential gains blocked by sampling/error; quantum-inspired classical useful in sparse cases
Schatzki et al.	Quantum-native data	NTangled dataset; QNN/QCNN tests	QNNs/QCNNs succeed on entangled states; avoid embedding/barren plateau issues
Baughman et al.	Explainability	Hybrid pipeline with classical explainers	Provides interpretability on small datasets; bridges feature an importance gap
Zaman et al.	Comprehensive survey	Hybrid VQC, kernels, taxonomy	Identifies NISQ limits, benchmarks and co-design needs
Qi et al.	NISQ enhancement	Tensor networks, pretrained models, RL, generative AI	TTN-VQC, BERT-QCNN enhanced performance; RL supports architecture search
Chen and Liang	Quantum arch. Search	Tutorial on VQC, RL/diff. Search; federated QML	QAS optimizes circuits; propose federated QML and model compression
Caro et al.	Generalization bounds	Learning theory + QCNN/QFT compiling	QML generalizes with polynomial/polylog data; few samples suffice for compiling
Thanasilp et al.	Trainability	Variance bounds, FI analysis, embedding sims	Dataset-induced barren plateaus found; encoding choice critical

[5] also uses real-world data sets analyzed with traditional feature importance methods to choose important features for quantum models.

Quantum Models and Algorithms: A broad range of QML models were investigated. Classical-to-quantum translation techniques used QSVM, Quantum k-means, and simple Quantum Neural Networks [1–3]. Hybrid quantum-classical models used

Variational Quantum Circuits (VQC) as the core backbone, sometimes coupled with classical neural networks, tensor networks, or pre-trained language models [6–8]. Quantum Convolutional Neural Networks (QCNNs) [4] and Quantum Long Short-Term Memory (QLSTM) [8] were used for sequence learning or feature extraction. Quantum kernels and hybrid optimization methods were also used to enhance the model's performance and scalability [6].

Training and Optimization Techniques: Training techniques aim at gradient-based optimization of parameterized circuits. Research on traditional data [3–10] examines generalization bounds and pinpointed trainability problems like barren plateaus and dataset-induced barren plateaus (DIBPs). Hybrid approaches [5–8] utilised traditional optimizers with quantum circuits, typically utilizing quantum architecture search (QAS) through reinforcement learning, evolutionary algorithms, or differentiable search to optimize circuit topologies. Preprocessing of data involves normalization, encoding to qubit states, and, where necessary, feature selection with classical explainability methods [5].

In total, the approaches capture a range of QML methods from purely quantum algorithms to hybrid structures optimized for NISQ machines, with focus on the balance between dataset selection, circuit construction, encoding schemes, and training paradigms. Such methodological differences enable the models to tackle various QML problems, such as generalization, trainability, noise tolerance, and real-world scalability (Table 2).

4 Discussions

The literature review shows that QML has developed along several complementary tracks. Traditional-data QML methods show that quantum algorithms such as QSVM and Quantum k-means can, in principle, be faster than their classical counterparts, but they are subject to major practical impediments such as state preparation, the restricted NISQ hardware capabilities, and error accumulation. These works emphasize that, absent effective encoding strategies and noise reduction measures, they remain difficult to access in practice due to encoding and noise challenges.

Hybrid quantum-classical approaches present a real-world solution to these challenges. With the use of quantum-native datasets [4] or classical explainers [5], researchers are able to enhance training efficiency as well as interpretability. Variational quantum circuits (VQCs) coupled with classical neural networks, tensor networks, or pre-trained language models [7, 8] further increase the expressiveness of QML models, whereas frameworks such as quantum architecture search (QAS) offer systematic methods to optimize circuit design within NISQ requirements. These efforts in aggregate suggest that the utilization of classical resources in conjunction with quantum models is critical for near-term application.

Theoretical studies [9, 10] provide key insights into the limitations and requirements of QML models. Generalization bounds show that even small datasets are enough for efficiently parameterized quantum circuits, as long as effective circuit complexity is managed [9]. On the other hand, the discovery of dataset-induced barren plateaus (DIBPs) [10] emphasizes that ill-designed encodings or strongly entangled inputs can prevent gradient-based optimization, restricting trainability. These results highlight the interaction between data structure, model architecture, and optimization strategy and indicate

Table 2. Overview of encoding strategies, architectures, and challenges in quantum machine learning studies.

Paper	Encoding strategy	Architecture/model used	Notes and challenges
Jadhav et al.	Amplitude and basis encoding for classical data	QSVM, Quantum k-means, QNNs	Speed-ups possible but limited by state preparation and noise
Abohashima et al.	Survey of angle, amplitude, and hybrid encodings	Pure quantum, quantum-inspired, hybrid models	Encoding central to classification accuracy; error mitigation critical
Wossnig	Amplitude encoding on classical datasets	Supervised QML models, complexity analysis	Measurement overhead and error scaling limit practical gains
Schatzki et al.	Entangled quantum-native datasets	QNNs, QCNNs	Avoids embedding overhead, mitigates barren plateau issues
Baughman et al.	Feature-selected encoding (classical preprocessing)	Small-scale QML models with explainability	Bridges feature importance gap; interpretable but limited scalability
Zaman et al.	Discusses multiple encodings (angle, basis, kernel methods)	Hybrid VQC, QML kernels	Highlights need for standardized benchmarks and hardware co-design
Qi et al.	Hybrid encodings with tensor networks	VQC + tensor networks, BERT-QCNN	RL and generative AI enhance circuit search; NISQ adaptation
Chen and Liang	Encoding for VQCs and federated learning	VQC, QLSTM, QAS frameworks	Automated circuit design via QAS; applications in federated QML
Caro et al.	Encodings for phase classification, QFT compiling	QCNNs, QFT-based circuits	Theoretical generalization bounds; few data samples can suffice
Thanasilp et al.	Encodings analyzed for barren plateaus	Variance bounds, Fisher Information methods	Dataset-induced barren plateaus identified; encoding choice critical

that effective deployment of QML demands careful co-design of circuits, encodings, and training procedures.

Generally, the presentation of this review reflects a clear path in QML research: from theoretical algorithmic accelerations to hybrid and application-driven approaches, ending with strict comprehension of generalization and trainability [9, 10]. Much progress has been made, but the problems of scaling models to real-world data, maintaining interpretability, and avoiding noise and barren plateaus persist. This intersection of theory, hybrid implementation, and practical limitations means that subsequent research needs to bring together architecture-aware, data-aware, and hardware-aware approaches in order to realize the full potential of QML.

5 Conclusion

This survey illustrates that Quantum Machine Learning (QML) evolves in various directions, ranging from theoretical models to experimental hybrid realizations on NISQ hardware. Clean quantum algorithms like QSVM, Quantum k-means, and QCNNs have the potential to provide speed-ups and improved feature representation, especially when used on quantum-native data. Quantum-classical hybrid methods then extend the applicability of QML to classical data even more while reducing hardware constraints, using methods like variational circuits, tensor networks, pre-trained classical models, and quantum architecture search. Even with these developments, primary challenges remain. Encoding of data, trainability problems such as barren plateaus and dataset-caused barren plateaus, and hardware constraints continue to limit scalability and performance. Guarantees of generalization are still limited, particularly in the case of large-scale classical datasets, calling for optimized circuit design and cautious hybrid integration.

In summary, QML research showcases a significant theoretical potential and growing practical achievements. The studies as a whole show that the most powerful strategy for harnessing QML in practice at present is to integrate quantum and classical resources by means of hybrid architectures, enhanced encoding techniques, and quantum architecture search. Scalable, trainable models and high-quality datasets are what upcoming work will need to concentrate on if QML is to reach its full potential.

References

1. Jadhav, A., Rasool, A., Gyanchandani, M.: Quantum machine learning: scope for real-world problems. In: Proceedings of International Conference Machine Learning and Data Engineering (2025)
2. Abohashima, Z., Elhoseny, M., Houssein, E.H., Mohamed, W.M.: Classification with Quantum Machine Learning: A Survey (2025)
3. Wossnig, L.P.: Quantum Machine Learning for Classical Data (A Dissertation) (2021)
4. Schatzki, L., Arrasmith, A., Coles, P.J., Cerezo, M.: Entangled Datasets for Quantum Machine Learning (2025)
5. Baughman, A., Yogaraj, K., Hebbar, R., Ghosh, S., Haq, R.U., Chhabra, Y.: Study of Feature Importance for Quantum Machine Learning Models. IBM (2025)
6. Zaman, K., Marchisio, A., Hanif, M.A., Shafique, M.: A Survey on Quantum Machine Learning: Basics, Current Trends, Challenges, Opportunities, and The Road Ahead (2025)

7. Qi, J., Yang, C.-H.H., Chen, S.Y.-C., Chen, P.-Y.: Quantum Machine Learning: An Interplay Between Quantum Computing and Machine Learning (2025)
8. Chen S.Y.-C., Liang, Z.: Introduction to Quantum Machine Learning and Quantum Architecture Search (2025)
9. Caro, M.C., Huang, H.-Y., Cerezo, M., Sharma, K., Sornborger, A., Cincio, L., Coles, P.J.: Generalization in Quantum Machine Learning from Few Training Data (2025)
10. Thanasilp, S., Wang, S., Nghiem, N.A., Coles, P., Cerezo, M.: Subtleties in the Trainability of Quantum Machine Learning Models (2025)
11. Maksimova, E.A., Lapina, M.A., Lapin, V.G., Rusakov, A.M.: Anthropomorphic model of states of subjects of critical information infrastructure under destructive influences. In: Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I. (eds.) Mathematics and its Applications in New Computer Systems. MANCS 2021. Lecture Notes in Networks and Systems, vol. 424. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97020-8_51

Assessment of the Level of Problem Disclosure in Technical Assignments Prepared Using Large Language Models

Karolina Pominova(✉) and Hao Hu

Samara National Research University Named After Academician S. P. Korolev, 34 Moskovskoye Shosse, Samara 443086, Russia
pominovak@bk.ru

Abstract. This paper evaluates the quality of technical specifications (TZ) generated by large language models (LLMs) for import substitution in civil aviation, focusing on automating systems in the Tu-204 aircraft to eliminate the flight engineer's role and mitigate human factor risks. Drawing on aviation standards such as GOST 34.602–2020, FAP-145, and ARP4754A, we develop a quantitative assessment methodology using a completeness coefficient K ($0 \leq K \leq 1$), which integrates keyword coverage (K_w), prompt quality (Kp), clarifying questions (K_q), and knowledge depth (K_r), weighted by $W = 0.4$, $P = 0.25$, $Q = 0.2$, $R = 0.15$. LLMs were prompted to identify engineer-responsible systems (e.g., power supply, engines, hydraulics, navigation) and formulate TZ responses, structured per industry templates. Analysis across subsystems yields an average $K \approx 0.58$, indicating moderate coverage ($K_w = 0.65$) but gaps in scenarios, metrics, and certification details ($K_p = 0.58$, $K_q = 0.5$, $K_r = 0.6$). Common LLM limitations include hallucinations, incomplete requirements, and non-compliance with norms like ASD-STE100. Cost estimates (e.g., 1.2M RUB for hydraulics vs. actual 4-5M) highlight verification needs. Results underscore LLMs' efficacy for draft generation, reducing labor by 20–30%, yet necessitating expert human-in-the-loop validation for safety-critical aviation projects. Future enhancements via fine-tuning and RAG integration could elevate K to 0.8 +, fostering AI-assisted documentation in high-stakes domains.

Keywords: Large language models (LLMs) · Technical specification (TZ) · Aviation automation · Tu-204 aircraft · Import substitution · Completeness coefficient K

1 Introduction

The Terms of Reference (ToR) is a document that defines the goals, requirements, conditions, and key characteristics of a project, product, or service. It serves as the foundation for development, agreement between the client and the contractor, and outcome control, clearly defining the requirements, conditions, and expectations for the project, product, or service. It also serves as the basis for quality control and work deadlines.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 517–526, 2027.
https://doi.org/10.1007/978-3-032-27834-0_48

In the aviation industry, ToR is the primary document that outlines the goals, requirements, and characteristics of the aircraft, system, or software being developed. Its main task is to provide a detailed description of the modified or newly created structure, its purpose, safety requirements, operational characteristics, as well as control and acceptance criteria to ensure compliance with all standards and regulations of the aviation industry.

The topic of the test Terms of Reference under consideration focuses on import substitution in the production of civil aircraft. The study formulated requirements for a series of systems in the Tu-204 aircraft project using AI.

When drafting the ToR, it is essential to rely on reliable and specialized sources of information: GOST standards, regulations, and specific documentation for civil aircraft as a template for its preparation. The drafting of the Terms of Reference in this industry was based on the idea of automating the Tu-204 aircraft by eliminating the role of the flight engineer as a risk to the entire system's operation and mitigating the impact of human factors in the aircraft management process.

The LLM was tasked with identifying the systems for which the flight engineer is responsible on the aircraft in question and providing answers to pre-prepared questions to assess the depth and quality of the material it presented. After receiving answers to all the questions related to the systems requiring automation, the LLM formatted them into the ToR, based on the documentation it deemed necessary.

This work aims to evaluate the quality of the ToR drafted with the assistance of AI. To assess the quality of the generated material, a special methodology has been developed, which involves the introduction of a specific coefficient. The value of this coefficient ranges from 0 to 1, depending on the level of topic coverage in the ToR. The closer the value is to 1, the more complete and executable the assignment; 0 characterizes a completely failed version of the ToR generated by AI.

The article explores the level of quality in Terms of Reference and project documentation that can be achieved through the use of AI.

2 Review of Existing Approaches

The development of technical documentation is a vast and complex task, requiring significant effort and high expertise from specialists. The question of automating the processes of creating technical specifications and project documentation has always attracted attention, and in recent years, there has been a surge of interest in this issue, driven by the emergence of artificial intelligence technologies and attempts to implement them.

An analysis was conducted of existing approaches to integrating AI into parts of the work of writing ToR and into various spheres of human activity.

Study [1] analyzes digital economy tools from the perspective of growth in key performance indicators affecting the efficiency of high-tech enterprises. It considers the role of artificial intelligence as a complement to human intelligence, opening a clear path for the development of innovative technologies, industrial robots, and the Internet of Things.

Articles [2, 3] note that traditional documentation processes suffer from problems of being static, quickly becoming outdated, and having high maintenance costs. Meanwhile, AI-driven dynamic documentation generation, using technologies such as natural

language processing, machine learning, and data mining, enables automatic content extraction from various sources like source code, user stories, etc., and generates documentation with real-time updates and contextually relevant content. This improves efficiency, reduces human errors, and enhances team collaboration by freeing up time.

Article [4] discusses the reduction in aircraft design cycle time due to the use of artificial intelligence. The research applies Fuzzy Logic (FL) and Neural Network (NN) methods. The Fuzzy Logic method involves four components and generates rules for approximating the WTS system of a light business jet.

Research [5] focuses on the integration of artificial intelligence (AI) into aviation automated systems to enhance the competitiveness of the United States in the aerospace industry. A decision matrix was developed to guide AI application, a formal methodology for assessing the quality enhancement of neural network systems was created, and the goal was set to transform AI from a novel technology into a standard component for developing advanced aviation systems.

Traditional aviation maintenance, based on scheduled inspections and reactive repairs, leads to suboptimal downtime and unplanned failures. The advancement of AI, Machine Learning (ML), and the Internet of Things (IoT) facilitates Condition-Based Maintenance strategies, detecting aircraft anomalies in real-time. A specialized platform with AI analytics was proposed to optimize processes and enhance system resilience [6].

Article [7] examines how modern Artificial Intelligence (AI) technologies can assist software developers in creating technical documentation that complies with the "AI Act". The authors demonstrate how developed tools can identify gaps in existing documentation according to the provisions of the "AI Act".

Work [8] presents AI-based Natural Language Processing (NLP) models that are trained for the contextual representation of technical documents containing text, figures, and tables. They are used for performing semantic search and extracting relevant data from large volumes of documents. By linking textual and non-textual data through an associative base, the developed semantic search system can provide more comprehensive answers in the context of user queries.

Work [9] presents a digital assistant platform designed for real-time interaction with experts, collecting procedural knowledge, and structuring it. The system operates through guided interaction where experts describe and demonstrate procedures, and the assistant extracts key information using structured dialogues, multimodal data processing, and contextual clarification. The collected data is processed by a Python-based server and integrated into a Unity-based interactive interface, allowing users to dynamically access documentation.

Work [10, 11] discusses the application of AI and Natural Language Processing methods for automatically generating accurate and context-dependent documentation for software code. This research highlights the potential of NLP-based tools for automating and standardizing documentation processes, offering a scalable solution to a long-standing challenge in software engineering.

Research [12] proposes enhancing API documentation using generative AI. In this work, generative AI is used to improve the accuracy, speed, and scalability of the generated API documentation. An automated API documentation generator was developed using natural language processing techniques applied to a large language model.

Work [13] notes that clinical documentation and data verification reduce the productivity of US physicians. An application based on artificial intelligence (AI) was proposed that could automate the process of creating medical records.

Work [14] points out that deep learning methods, along with automatic metrics, are used for the automatic generation of code documentation. During the research, three state-of-the-art approaches were replicated, and the documentation itself was evaluated against three parameters—language, content, and efficiency—using automatic metrics.

3 Key Principles for Evaluation

The Terms of Reference is a document that defines the key characteristics of the object under development. A ToR is considered sufficient if project documentation can be developed based on it without the need for additional fundamental clarifications. Any additional request regarding the ToR lowers its quality assessment. The criterion for the sufficiency of information from the ToR for developing project documentation is determined by the completeness and unambiguity of the information required for subsequent design. The main objective of the model is to develop a quantitative assessment system specifically designed to determine the quality of the ToR [15]. The study introduces a special coefficient that ranges from 0 to 1. This coefficient allows for judging the feasibility of successfully implementing the ToR. A value of 1 is assigned to an ideal ToR, indicating that the information from the ToR is sufficient for successfully developing the technical documentation, and no additional information will be required during the work. A score of zero corresponds to a complete failure of the implementation under consideration. However, these are ideal cases that are not realized in practice. In real life, the assessment will be intermediate, describing the depth of topic coverage. This value can be assessed based on the degree of necessity for additional information to write the ToR. The more questions and clarifications required during the preparation of project documentation, the lower the cumulative score.

To facilitate quality assessment, it is useful to divide the coefficient values into gradations and introduce fractional values. These gradations will help assess the quality level of the ToR in more detail. The criterion for the sufficiency of information from the ToR for developing project documentation is determined by the completeness and unambiguity of the information required for design. The following approach is proposed for assessing the value of the coefficient K. In the first stage, a list of keywords related to the topic of the ToR is generated. This list is formed by querying an LLM. Next, based on the list of keywords, a set of phrases for queries to the LLM is formulated. The answers to these queries constitute the content of the ToR. The completeness of the ToR content and the depth of detail of the information from the ToR depend on the completeness of the set of phrases. It is necessary to develop an appropriate procedure for assessing the completeness and depth of topic coverage.

The levels of problem disclosure in ToR depend on the context, the document's objectives, and the development stage. Typically, several levels of detail are distinguished, which determine how deeply and thoroughly the problem is described. Here are the main levels of problem disclosure in a ToR:

1. General Level (Conceptual): The problem is described in general terms, without technical details. It specifies what exactly is not working or requires improvement and what goals need to be achieved.
2. Functional Level: The problem is disclosed through specific functions that are not performed or require refinement. Key user scenarios or functional shortcomings are indicated.
3. Technical Level: The problem is described considering technical aspects, including system architecture, technology limitations, integrations, etc. Specific technical causes and possible solutions are indicated.
4. Detailed Level (Scenarios and Metrics): The problem is disclosed through specific usage scenarios, performance metrics, and expected outcomes. Numerical indicators, SLAs (Service Level Agreements), and acceptance criteria are specified.

The more specific measurable parameters, technical specifications, error handling options, examples, and visualizations a ToR contains, the higher its level of detail. An ideal ToR should allow the developer to start work without additional clarifications [16, 17].

As of the current date (2025), fully automated systems for assessing the depth of a ToR do not exist, but there are tools and methods that help partially automate this process.

4 Mathematical Modeling

The primary objective of the mathematical model is to develop a quantitative assessment system specifically designed to measure the completeness of technical requirements documentation for projects implementing a large language model. The main output of this system is the completeness coefficient K, which ranges from *0 to 1* ($0 \leq K \leq 1$). A higher K value indicates more comprehensive requirements documentation.

There are three main categories of evaluation parameters:

1. Keywords;
2. The set of phrases for the large language model necessary to cover the topic;
3. The presence of clarifying questions.

The probability of errors in a ToR compiled using AI depends on numerous factors: the quality of the source data, the complexity of the task, the model configuration, and post-verification. Discrepancies can be identified during the document structure review against commonly accepted standards like GOST 34 or IEEE 830, often chosen as benchmarks; during the analysis of requirement formulations for their correctness; during acceptance checking via measurable indicators; and through the use of auxiliary materials such as diagrams, tables, reference processes to regulations, etc. [18, 19].

The quantitative assessment system for the completeness of Technical Requirements (TR) documentation for an LLM project is presented below:

$$\mathrm{K} = \left(\mathrm{W} \cdot \mathrm{K_w} + \mathrm{P} \cdot \mathrm{K_p} + \mathrm{Q} \cdot \mathrm{K_q} + \mathrm{R} \cdot \mathrm{K_r}\right)$$

where: K is the final completeness coefficient; K_w is the keyword coverage coefficient ($0 \leq K_w \leq 1$); K_p is the LLM phrase set quality coefficient $\left(0 \leq K_p \leq 1\right)$; K_q is the

clarifying questions elaboration coefficient $\left(0 \leq K_q \leq 1\right)$; K_r is the LLM knowledge depth coefficient.

W, P, Q, R are weighting coefficients reflecting the importance of each category. The sum of the weighting coefficients must equal 1, i.e.,

$$\mathrm{W} + \mathrm{P} + \mathrm{Q} + \mathrm{R} = 1$$

5 Evaluation Dimensions

5.1 Keywords

The formula for calculating the keyword coverage ratio:

For the keyword set, it is necessary to evaluate how well the generated documentation covers all critical aspects of the LLM project. Using expert analysis of the obtained data to check for the presence of a predefined set of key topics, each keyword is evaluated in a binary manner (0 or 1) based on the principle of "mentioned/covered."

The formula for calculating the keyword coverage coefficient is:

$$\mathrm{K_w} = \frac{\mathrm{L}}{\mathrm{M}}$$

where: L is the sum of the weights of the keywords found; M is the total sum of the weights of all keywords in the list.

5.2 Set of Phrases for the Large Language Model

At the stage of assessing the quality of the ToR based on the set of phrases created using LLMs, it is necessary to develop a set of queries that will cover the key aspects of the project, divided into categories such as functional requirements, technical specifications, industry standards, and operational scenarios.

An example prompt for automating the Tu-204 systems: "Describe the functions of the PNK-204 for eliminating the flight engineer, including integration with FADEC and GLONASS." The response obtained is compared against a benchmark, which includes a list of functions (e.g., engine monitoring, power management), technical specifications (e.g., memory capacity, power requirements, compatibility with onboard systems), references to standards (e.g., FAP-145, GOST 34.602–2020), and usage scenarios (e.g., automating navigation during flight with a specified course accuracy).

The success criterion is manually defined—for example, the response must include at least three functions, one reference to a standard, and a timeframe for execution. The LLM phrase set quality coefficient is calculated as the proportion of prompts with complete responses:

$$\mathrm{K_p} = \frac{\mathrm{S}}{\mathrm{T}}$$

where: S is the number of successful responses; T is the total number of prompts.

The presence of a detailed set of phrases with benchmarks in the documentation directly influences the evaluation, demonstrating that the team has considered not only "what" the model should do but also "how" it should behave in non-standard situations.

5.3 Presence of Clarifying Questions

To assess the depth of requirement elaboration, it is necessary to reduce the level of uncertainty using the Clarifying Questions Elaboration Coefficient. The methodology involves an expert, who is well-versed in the subject area and LLM, compiling a list of 10–15 clarifying questions to ensure consistency with the documentation regarding key aspects, similar to the set of phrases. The coefficient is based on the proportion of questions for which the required answers can be obtained within the generated documentation.

The Clarifying Questions Elaboration Coefficient is calculated using the formula:

$$K_q = 1 - \left(\frac{N_{unknown}}{N_{total} \cdot 0.8} \right)$$

where: $N_{total} \cdot 0.8$ is the total number of clarifying questions generated by the expert, adjusted by a correction factor to reduce the weight of minor questions; $N_{unknown}$ is the number of questions for which no clear answer can be obtained within the generated documentation.

5.4 Knowledge Depth

To evaluate the depth of technical documentation comprehension, the Knowledge Depth Coefficient is used, which determines how thoroughly the technical aspects of the project are covered. This parameter reflects the ability of the ToR to provide sufficient and comprehensive information about systems, compatibility, limitations, etc., minimizing the need for additional clarifications. This coefficient is calculated based on a checklist developed by aviation industry specialists, which considers key technical details required for the development of automated systems and their implementation in aircraft. It is calculated using the formula:

$$K_r = \frac{Y_{unknown}}{Y_{total}}$$

where: $Y_{unknown}$ is the total score of the checklist items; Y_{total} is the maximum possible score.

This coefficient enhances accuracy for complex aviation ToRs and improves detail, simplifying post-verification by industry specialists.

6 Analysis of the Results

In aircraft manufacturing, the application of artificial intelligence for creating technical specifications opens new possibilities, enabling the optimization of complex system designs, such as aircraft and their components. For the Tu-204 aircraft, where the focus of automation is on eliminating the role of the flight engineer to minimize the human factor as an unstable variable, the use of AI is highly relevant.

The structure of the generated Technical Specification (TS) was analyzed both for individual systems under the flight engineer's responsibility and comprehensively in

terms of knowledge depth, completeness, and reliability. This was done using the completeness coefficient K, linked to the parameters of the primary formula with desired indicators. Costs were calculated in rubles at the average exchange rate of summer 2025 (approximately 80 rubles per US dollar, based on June-August data). The coefficient values were also determined: K_w (keywords, $W = 0.4$), K_p (query quality, $P = 0.25$), K_q (clarifying questions, $Q = 0.2$), and K_r (knowledge depth, $R = 0.15$). Based on the analysis of the AI response model across systems (power supply, engines, hydraulics, navigation, communication, flight control, air conditioning, safety), the average $K \approx 0.58$, indicating a need for refinement in 42% of cases to comply with FAP-145 and GOST 34.602–2020 standards [20, 21].

7 Overall Conclusion

The overall TS received a $K \approx 0.58$ due to moderate completeness ($K_w = 0.65$, covering key terms like "FADEC" but not all standards), limited query quality ($K_p = 0.58$, lacking scenarios and metrics), partial elaboration of questions ($K_q = 0.5$, insufficient clarifications per system), and average depth ($K_r = 0.6$, basic parameters are present but lack complex details). The weights, determined through Pearson correlation and gradient descent, emphasize the priority of completeness ($W = 0.4$), but gaps in details and reliability reduce the final score. AI is effective for drafting but requires expert verification to meet aviation requirements, especially regarding costs (e.g., 1.2 million rubles for hydraulics versus the actual 4–5 million) and certification.

For instance, when planning the replacement of the power supply system, AI proposes integrating PNK-204 for monitoring generators and batteries but overlooks certification nuances, leading to 2–3 clarifying questions from experts and reducing completeness to an average level. Similarly, in control and engine systems, the focus is on implementing FADEC for automatic parameter adjustment, but reliability drops due to the absence of real-world cases, such as the PS-90A upgrade in the Tu-204SM. In communication and flight control systems, proposals include FMS from Rockwell Collins, with scores of 7.33, but lack integration risks, necessitating statistical calibration of K weights via Pearson correlation for better alignment with aviation norms. For air conditioning, heating, and safety, the focus is on HUMS diagnostics, where replacement costs 1.8 million rubles with a 20% saving on maintenance, compared to 0.7 million for modifications. However, the 15-year long-term service life makes the former more cost-effective, underscoring the need for expert verification to minimize errors in the TS [22].

Based on the analysis of various subsystems, from power supply to safety, it is evident that AI can generate basic TS structures, but they often lack sufficient depth. Testing of AI responses confirms that with K values between 0.55 and 0.65, documents require refinement in most cases, particularly to comply with FAP and GOST standards, where question normalization $1/(1 + 0,5 \cdot N)$ and gradient descent for weights ensure accuracy. The overall approach with Pearson correlation ensures the model's adaptability to real-world projects in aircraft manufacturing, where balancing automation and expertise is key to success.

This highlights the balance between automation and human oversight in aviation documentation, where an in-depth analysis of Tu-204 systems demonstrates the potential

for cost and risk reduction, but only through a comprehensive assessment of completeness for practical implementation [23, 24].

8 Conclusions and Summary

AI can generate ToRs of acceptable quality, but human expertise remains essential for critical systems. The optimal approach is to use AI as an assistant for initial drafting, followed by multi-level verification, thereby simplifying the labor-intensive nature of the process.

References

1. Tikhonov, A.I., Sazonov, A.A., Kuzmina-Merlino, I.: Digital production and artificial intelligence in the aircraft industry. Russ. Eng. Res. **42**(4), 412–415 (2022)
2. Ajeigbe, K., Emma, O.: Dynamic Documentation Generation with AI (2024)
3. Hull, E. et al.: Requirements Engineering [electronic resource]
4. Oroumieh, M.A.A., Malaek, S.M.B., Ashrafizaadeh, M., et al.: Aircraft design cycle time reduction using artificial intelligence. Aerosp. Sci. Technol. **26**(1), 244–258 (2013)
5. Broderick, R.: Knowledge-based Aircraft Automation: Managers Guide on the Use of Artificial Intelligence for Aircraft Automation and Verification and Validation Approach for a Neural-based Flight Controller. NASA-CR-205078 (1997)
6. Kabashkin, I., Perekrestov, V., Pivovar, M.: AI-driven fault detection and maintenance optimization for aviation technical support systems. Processes **13**(3), 666 (2025)
7. Sovrano, F., et al.: Simplifying software compliance: AI technologies in drafting technical documentation for the AI Act. Empir. Softw. Eng. **30**(3), 91 (2025)
8. Gammack, J., et al.: Semantic knowledge management system for design documentation with heterogeneous data using machine learning. Procedia CIRP **109**, 95–100 (2022)
9. Cappello, L.: An AI-powered digital assistant framework for automated procedure documentation (2023)
10. Ibrahim, A., et al.: Using NLP and AI to enhance software documentation and code comprehension. Spectr. Eng. Sci. **3**(5), 195–204 (2025)
11. Pohl, K.: Requirements engineering fundamentals: a study guide for the certified professional for requirements engineering exam-foundation level-IREB compliant. Rocky Nook, Inc. (2016)
12. Dhyani, P. et al.: Automated API docs generator using generative AI. In: 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) , pp. 1–6. IEEE (2024)
13. Lin, S.Y., Shanafelt, T.D., Asch, S.M.: Reimagining clinical documentation with artificial intelligence. Mayo Clinic Proc. **93**(5), 563–565 (2018)
14. Hu, X. et al.: Correlating automated and human evaluation of code documentation generation quality. ACM Trans. Softw. Eng. Methodol. (TOSEM) **31**(4), 1–28 (2022)
15. Ivanov I.I.: Terms of Reference
16. ARP4754A S. A. E.: Guidelines for development of civil aircraft and systems. SAE International (2010)
17. Gledhill, C.: Simplified technical English: from the description of a new technolect to a critique of New New-speak, 32eme Colloque du GERAS (2011)
18. Leveson N. G. Engineering a safer world: Systems thinking applied to safety. The MIT Press, 2016, 560

19. Johnson, L.A. et al.: DO-178B: Software considerations in airborne systems and equipment certification **199**, 11–20 (1998)
20. Easa, C.: Certification specifications for large aeroplanes, cs 25 (2009)
21. Evaluating Text Generation with BERT. In: International Conference on Learning Representations (ICLR)
22. Retzlaff, C.O., et al.: Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. J. Artif. Intell. Res. **79**, 359–415 (2024)
23. Young, T.M.: Performance of the jet transport airplane: analysis methods, flight operations, and regulations. John Wiley & Sons (2017)
24. Dmitriev, V.G., Samokhin, V.F.: Equivalent method of acoustic certification of aircraft family with the modified power plant. AGI Sci. J. 48(2) (2017)

BELM: An AI-Based Hybrid Model for Detecting and Mitigating Cyberbullying Against Celebrities on Social Media

Boppuru Rudra Prathap[1(✉)], Vandana Reddy[2], and Maxim Medvedev[3]

[1] M. S. Ramaiah University of Applied Sciences, Bangalore, India
brprathap@gmail.com
[2] CHRIST Deemed to Be University, Bangalore, India
[3] North Caucasus Federal University, Stavropol, Russia

Abstract. The detection and mitigation of cyberbullying largely depends on the use of Artificial Intelligence (AI) and Machine Learning (ML). By utilizing these technologies, it is possible to analyze extensive data sets for signs of bullying tendencies that enable timely intervention. This study suggests an exceptional mixed method that utilizes BERT (Bidirectional Encoder Representations from Transformers) BERT's contextual understanding with LSTM (Long Short-Term Memory) networks' sequence processing capabilities in detecting and mitigating cyberbullying against celebrity Figs on social media. More accurate, precise, recalled, and F1 score than those found in the individual models can be achieved by integrating the two powerful models. The BERT model has accuracy rate of 89%, precision rate 87%, recall rate 86% and F1-score at 86.5% while the LSTM technique has an accuracy level of 85%, precision level of 84%, recall level of 83% and F1-score at 83.5%. The research carried out has used this hybrid model which was trained and tested on social media posts showing that it is applicable in real-world scenarios with accuracy of 92.8%, precision of 92.3%, recall of 91.5%, and F-1-Score is equal to 93%.

Keywords: Bidirectional encoder representations from transformers- BERT · Long short-term memory- LSTM · Cyberbullying · Celebrity trolls · Artificial intelligence · Machine learning

1 Introduction

Cyberbullying is when a person is harassed, threatened, or demeaned through the use of digital platforms. Bullying could happen on social media, forums, and other online communities, which might result in huge emotional and psychological damage to a victim. Celebrities often suffer from cyberbullying since they are open to the public. This relentless trolling can result in mental health problems, loss of public image, and withdrawal from public life in extreme cases. Artificial intelligence and machine learning have been instrumental in identifying and mitigating instances of cyberbullying. Such technologies

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 527–544, 2027.
https://doi.org/10.1007/978-3-032-27834-0_49

would be able to analyze large volumes of data for patterns indicative of bullying behavior that would allow timely interventions. The study compares the effectiveness of the two latest AI/ML models developed toward the identification and removal of cyberbullying targeted at celebrities. In this regard, the research will base its comparisons on the accuracy of detection, precision, recall, and real-world applicability. In the context of cyberbullying, it has been one of the major problems in India, especially with respect to Bollywood celebrities, ever since the evolution of social media. Bollywood celebrities are always trolled, harassed, bullied relentlessly online. The following is an overview of the current landscape, backed by available statistics and research findings. In most cases, Bollywood celebrities have often used social media platforms like Twitter, Instagram, and Facebook to reach out to their fans. Since this makes them very visible, they are also the prime targets of cyberbullying. Types of Cyberbullying are: i. Trolling which is negative, inflammatory comments intended to provoke and upset, ii. Harassment which is repeated, targeted abuse aimed causing distress, iii. Defamation which is spreading false information to damage reputation and iv. Threats that are Direct threats of violence or harm. It is depressing to learn the key statistics of the same where a study by the Internet and Mobile Association of India revealed that close to 80% of celebrities in India have faced some form of cyberbullying. Twitter and Instagram are the two major platforms where Bollywood celebrities face the most cyberbullying. Female celebrities face a higher rate of cyberbullying as compared to their male counterparts. A leading Indian news agency reported from a survey conducted in 2022 that 70% of female Bollywood celebrities have been harassed online severely. High-profile cases of cyberbullying occur very frequently, where the instances reported rise to the headlines every successive year. The relentless cyberbullying can pave ways to critical mental health problems, resulting in anxiety and depression, and even the most serious instance: suicidal thoughts. People like celebrities such as Deepika Padukone have come out to raise awareness about mental health and the bruises that cyberbullying inflicts. Cyberbullying also tends to negatively affect the professional lives of celebrities as well. Defamatory online campaigns dented their public image, causing them a loss of endorsements and work opportunities. India has under the Information Technology Act, 2000, and IPC enabling laws for combating cyberbullying and other forms of harassment online. The law is often turned to by celebrities to counter stronger cases of cyberbullying. Cyberbullying against the Bollywood celebrity fraternity in India has happened in several ways, including mental and professional life. Some mitigating steps are being taken, though—stronger legal frameworks, solutions driven by the platform, raising public awareness—are all needed if cyberbullying is to be minimized. BERT is a state-of-the-art model in natural language processing, particularly adept at contextual understanding in text.

One of the main reasons is that BERT's architecture will be able to look at the full sentence on either side, allowing it to do a very fine job of picking up subtle nuances in language. On application to cyberbullying detection, BERT identifies complex bullying patterns and contextual cues which may be skipped by simpler models. That is what turns it potentially useful for detecting subtle, context-dependent instances of cyberbullying targeted at celebrities. One of the many types of recurrent neural networks, LSTMs are capable of learning long-term dependencies within sequential data. LSTMs are therefore very good at processing and analyzing sequences of text, making them very good at

understanding the flow of conversations or comment threads. In this context of cyberbullying detection, LSTMs can recognize bullying behavior during extended interactions; that is, picking out persistent trolls depends on being able to identify bullying over a long enough period. This is helpful in recognizing patterns of harassment across several posts or messages, due to the ability of the LSTM to remember and use information across long sequences. Hybrid models can then be built that meld the technical capabilities of the context understanding of BERT with the sequential processing capabilities of the LSTM to significantly reduce cyberbullying and celebrity trolling. These models are used to analyze huge amounts of social media data in real-time, providing timely detection and interventions to help protect celebrities from online abuses, which formed the research work carried out.

Further the paper is organized into sections where Sect. 2 speaks about the literature review done for understating and identifying the solution for the same. Section 3 is about the methodology adopted to carry out the research. Section 4 deals with AI and ML methods applied to get the solution. Section 5 presents the results and discussion followed by conclusion and references.

2 Literature Review

Development in Artificial Intelligence and Machine Learning-based applications for the detection and mitigation of cyberbullying has been at par for the past years. BERT and LSTM networks are among the most efficient models. Zhong et al. [1] revealed that BERT works well when it comes to the identification of subtle bullying contexts related to social media posts, able to attain high accuracy due to its bidirectional nature. Wang et al. [2] demonstrated that fine-tuning BERT using a dataset containing subtle trolling behaviors, specifically celebrity tweets, significantly enhanced detection. Gupta and Kumar [3] examined the applicability of multilingual BERT in cross lingual cyberbullying detection and concluded that it had proved versatile with comparative efficiency. Patel et al. [4] introduced attention mechanisms into BERT and improved cyberbullying detection in long comment threads. Chen and Li [5], proposed a hybrid model that integrated the strengths of BERT with other feature extraction methods to improve both precision and recall in cyberbullying detection. Khan and Rizvi [6], employed an LSTM network for detecting sequential patterns in text indicative of bullying, achieving notable improvements in recall. Nguyen and Le [7] have pointed out some of the features of LSTM about handling long conversations that may find applicability in bullying detection on comment threads and social media interactions. Park and Lee [8] proposed to integrate it with sentiment analysis to explain the emotional context of possible bullying for improving accuracy in its detection. Zhang et al. [9] integrated the attention mechanism into LSTM networks to enhance the ability of focusing on the text parts relevant for assuring more accurate cyberbullying detection. Kim and Kwon [10] put forward a hierarchical LSTM model to handle text at multiple levels, from sentence-level to document-level bullying patterns. A study by Sharma et al. (2023) integrated BERT for context understanding and LSTM for sequential data processing and delivered superior results in detecting cyberbullying against celebrities. Verma and Singh (2023) have presented a real-time cyberbullying detection system using a hybrid BERT-LSTM model that has

high accuracy and efficiency. Lee et al. [13] evaluated these hybrid models on grounds of scalability. The findings of the test were that the integration of BERT with LSTM can support large datasets. Kumar et al. [14] integrated emotion recognition techniques with BERT-LSTM models for the improved identification of emotionally expressive bullying instances. Ahmed and Raj [15] applied these hybrid models to several social media platforms, proving to be feasible in varying contexts. Joshi and Desai [16], used a comparative study between BERT and LSTM. It has been observed that in this study, for contextual understanding, the former had the upper hand, while the latter is more efficient for sequence processing. Another work of Saha and Roy [17], used standalones against hybrid models of BERT and LSTM and has observed that hybrid models turn out to be the ones which yield the best overall performance. Malhotra et al. [18], experimented with transfer learning on BERT and LSTM and have shown that the pre-trained models adapt really well to the task at hand—cyberbullying detection. Srivastava and Bansal (2023) stressed the necessity of finetuning models on domain-specific data to help them become more relevant and accurate in celebrity trolling detection. Pandey and Sharma (2022) discussed some ethical dilemmas raised by the deployment of AI models for cyberbullying detection, showing their transparency and fairness. Reddy et al. (2023) worked on BERT and LSTM models' explainability in order to find out how they make decisions in the process of detecting cyberbullying. Jain and Mehta [22] have applied cyberbullying detection techniques to related areas such as hate speech and misinformation, thus showing the versatility of the BERT and LSTM models. Singh and Gupta [23] have further worked on adversarial training methods in order to strengthen BERT and LSTM models against such attacks, thus making them more reliable in their results. Thakur et al. [24] contributed continuous learning frameworks of BERT and LSTM models to accommodate the evolving patterns of bullying in real time. Bose and Rao [25] integrated text, image, and video analysis with BERT and LSTM to give an end-to-end solution for cyberbullying detection in multimedia platforms. From this literature review, it can be seen that the area involving the use of BERT and LSTM models for cyberbullying detection has undergone tremendous development. Of the two models, the hybrid approach has been very promising in addressing such complexities of cyberbullying and celebrity trolling. Future research should therefore continue to fine-tune these models with focuses on scalability, real-time detection, and ethical deployment. The Table 1 below provides a summation of the research gaps and limitations, tools used, and accuracy results in BERT and LSTM methods for cyberbullying detection.

2.1 Key Identifications

i. Research Gaps: Common gaps include the need for real-time applications, domain-specific fine-tuning, multilingual support, and effective cross-platform detection.
ii. Limitations: High computational costs, scalability issues, the requirement for large labeled datasets, and the complexity of hybrid models are significant limitations.
iii. Tools Used: Python, TensorFlow, PyTorch, Keras, BERT, LSTM, Attention mechanisms, Sentiment Analysis.
iv. Accuracy Results: Accuracy ranges from 88.9% to 94.0%, with hybrid BERT-LSTM models generally achieving higher accuracy.

Table 1. Research gaps, limitations, tools used, and Accuracy Results in BERT and LSTM Methods for Cyberbullying Detection.

Study	Research gaps	Limitations	Tools used	Accuracy (%)
Zhong et al. (2021)	Limited real-time application	High computational cost	Python, TensorFlow, BERT	91.5
Wang et al. (2022)	Need for domain-specific models	Scalability issues	Python, PyTorch, BERT	92.3
Gupta and Kumar (2021)	Multilingual support	Lack of cross-platform testing	Python, Keras, Multilingual BERT	90.1
Patel et al. (2023)	Handling long comment threads	Requires large labeled datasets	Python, TensorFlow, Attention-BERT	93.2
Chen and Li (2022)	Integration with other data types	Complexity in hybrid models	Python, PyTorch, BERT, Feature Extraction	91.8
Khan and Rizvi (2021)	Long-term pattern analysis	Requires sequential data preprocessing	Python, Keras, LSTM	89.7
Nguyen and Le (2021)	Extended conversation analysis	Language-specific limitations	Python, TensorFlow, LSTM	88.9
Park and Lee (2021)	Emotional context understanding	Integration with other emotional models	Python, PyTorch, LSTM, Sentiment Analysis	90.5
Zhang et al. (2022)	Attention mechanism optimization	Limited scalability	Python, TensorFlow, Attention-LSTM	91.0
Kim and Kwon (2022)	Multi-level text processing	Requires extensive training time	Python, Keras, Hierarchical LSTM	92.0
Sharma et al. (2023)	Real-time detection enhancement	Combining model complexities	Python, TensorFlow, BERT, LSTM	94.0
Verma and Singh (2023)	Real-time implementation	High resource requirements	Python, TensorFlow, BERT-LSTM Hybrid	93.5
Lee et al. (2024)	Scalability and performance testing	Resource-intensive training	Python, PyTorch, BERT-LSTM Hybrid	92.7

(*continued*)

Table 1. (*continued*)

Study	Research gaps	Limitations	Tools used	Accuracy (%)
Kumar et al. (2023)	Emotion recognition integration	Complex model interpretation	Python, TensorFlow, BERT-LSTM Hybrid	93.0
Ahmed and Raj (2022)	Cross-platform detection	Platform-specific optimizations	Python, Keras, BERT-LSTM Hybrid	91.4
Joshi and Desai (2021)	Comparative analysis depth	Limited dataset sizes	Python, TensorFlow, BERT, LSTM	90.8
Saha and Roy (2022)	Hybrid vs. standalone models	Model training complexity	Python, PyTorch, BERT, LSTM	92.6
Malhotra et al. (2023)	Transfer learning effectiveness	Adaptation to new data types	Python, TensorFlow, BERT, LSTM	91.9
Srivastava and Bansal (2023)	Domain-specific fine-tuning	High customization required	Python, Keras, BERT, LSTM	92.2
Pandey and Sharma (2022)	Ethical implications	Transparency in model decisions	Python, TensorFlow, BERT, LSTM	89.5
Reddy et al. (2023)	Explainable AI integration	Complexity in explanations	Python, PyTorch, BERT, LSTM	90.7
Jain and Mehta (2023)	Cross-domain applications	Model generalizability	Python, TensorFlow, BERT, LSTM	91.3
Singh and Gupta (2024)	Adversarial training	Robustness against attacks	Python, Keras, BERT, LSTM	91.6
Thakur et al. (2024)	Continuous learning frameworks	Model adaptability	Python, PyTorch, BERT, LSTM	92.4
Bose and Rao (2023)	Multi-modal analysis	Integration complexity	Python, TensorFlow, BERT, LSTM	93.3

3 Methodology

The Hybrid BERT-LSTM model is built on the individual strengths of BERT and LSTM networks to enhance cyberbullying detections, focusing significantly on those cases of targeted celebrity trolling. Therefore, the approach proposed here actually underlines

the fusion of BERT's contextual understandings with LSTM's capability of capturing sequential dependencies into a powerful system for the detection of forms of complex cyberbullying. It collects data from different social media platforms like Twitter, Instagram, and Facebook. This is directed toward posts and comments that are targeted at celebrities. Data annotation for cyberbullying content is done by experts, with special labels for insults, threats, and trolling. The data preprocessing uses tokenization and is processed further with padding and truncation, together with a special token. Tokenization: Text data is tokenized using a BERT tokenizer by dividing it into subword units. Padding and Truncation: Sentences are then padded to a unified length or more and also truncated into the required model input size. Special tokens include attributes added to the text sequences, such as [CLS] for the beginning of a sentence and [SEP] as a separator token.

The model consists of a BERT encoder, where a pre-trained BERT encoder is used in order to obtain contextual embeddings from the input text. Further fine-tuning of BERT on the annotated dataset makes it adapt to the specific context required for performing cyberbullying detection in celebrity interactions. In the LSTM layer, output embeddings from BERT are injected into the LSTM layer so that the model can capture sequential dependencies within the text; this will show how bullying language flows and develops. This gets enhanced with an attention mechanism that allows the LSTM to focus on relevant parts of the sequence in order to put up subtle cues of bullying. The classification layer: Fully connected layers analyze the final representations from the LSTM to perform classification. For the output layer, a softmax function is applied to yield probabilities for classes of cyberbullying.

The training and evaluation take place with the loss function, where the model is subjected to be trained by the cross-entropy loss function adopted for multi-class classification problems. In this method, Adam optimization is used for adaptive learning rate in achieving good convergence. The training of the model with a specified number of epochs is then executed, setting the batch size according to computational efficiency and performance, which serves as the compromise. The assessment of performance could be based on a number of metrics including, but not restricted to, accuracy, precision, recall, and F1-score. Cross-validation techniques are utilized to ensure the model developed is robust over different subsets of the data. A confusion matrix is created for the performance evaluation with respect to the model in correctly classifying various forms of cyberbullying. The mitigation strategies, in real-time monitoring with the model integrated into social media platforms, monitor posts and comments in real-time. The content is flagged and sent for further review by human moderators in case it is suspicious. Automated responses are generated to warn or block users engaging in bullying behavior. Awareness messages regarding cyberbullying are sent across to the offenders. Notifications are sent to the obligated celebrities to support them with options where they can seek help. Mechanisms for reporting and blocking abusive users are available on the platform. Figure 1 show the brief summary of the pre-trained BERT-LSTM hybrid Proposed architecture and it is having following input/output and step1 to step 10 procedure.

INPUT*:* Social media Tweet.

OUTPUT: The probability of Cyberbullying content and user cumulative score on Cyberbullying.

STEP 1 Data Collection:

- **Source:** Social media posts, forums, and online communities are monitored for user interactions such as likes, dislikes, and comments.
- **Purpose:** Collects raw text data from various user interactions across these platforms to serve as input for the LSTM network.

STEP 2 Feature Extraction:

- **Processing:** The collected text data undergoes feature extraction to highlight relevant characteristics that may indicate online misconduct.
- **Significance:** These features include elements like sentiment, keyword usage, and context, which help in analyzing the content.

STEP 3 LSTM Network Analysis

- **LSTM Network Role:** The LSTM (Long Short-Term Memory) network processes the extracted features, analyzing the sequence of user interactions to detect potential online misconduct.
- **Contextual Awareness:** The LSTM considers the context of the interactions, enabling it to recognize patterns of misconduct over time.

STEP 4 Evaluation Metrics

- **Assessment:** The LSTM network ranks detected instances of online misconduct based on predefined evaluation metrics.
- **Severity Determination:** This ranking helps in determining the severity of the misconduct and the necessary course of action.

STEP 5 Machine Learning Algorithm

- **Security Group:** This algorithm further processes the ranked instances by rating and grading the comments' context using pre-trained models.
- **Context Grading:** It assesses the context and assigns a score to the misconduct, which influences subsequent actions.

STEP 6 Legal and Regulatory Compliance

- **Compliance Checks:** The platform ensures that actions taken align with legal and regulatory frameworks, safeguarding against wrongful account actions.
- **Ethical Considerations**: In addition to legal compliance, the platform considers ethical aspects, ensuring fair treatment of users.

STEP 7 User Feedback and Iteration.

- **Feedback Loop:** User feedback is collected and integrated into the system, allowing for continuous improvement of the LSTM network's detection capabilities.

- **Adversarial Example Handling:** The system is also equipped to handle adversarial examples—deliberate attempts to bypass detection—by iterating and refining its algorithms.

STEP 8 Decision Making

- Severity Score Threshold: Based on the severity score derived from the evaluation and grading process, the system decides whether to freeze the user account or initiate automated deletion.
- Cumulative Score Evaluation: It checks whether the cumulative score of detected misconduct exceeds the threshold for serious action.

STEP 9 Automated Actions

- User Account Management: If the severity is high, the system can automatically freeze or delete the user account.
- Escalation Protocol: For less severe cases, an escalation protocol may be followed, which involves monitoring the user more closely rather than immediate action.

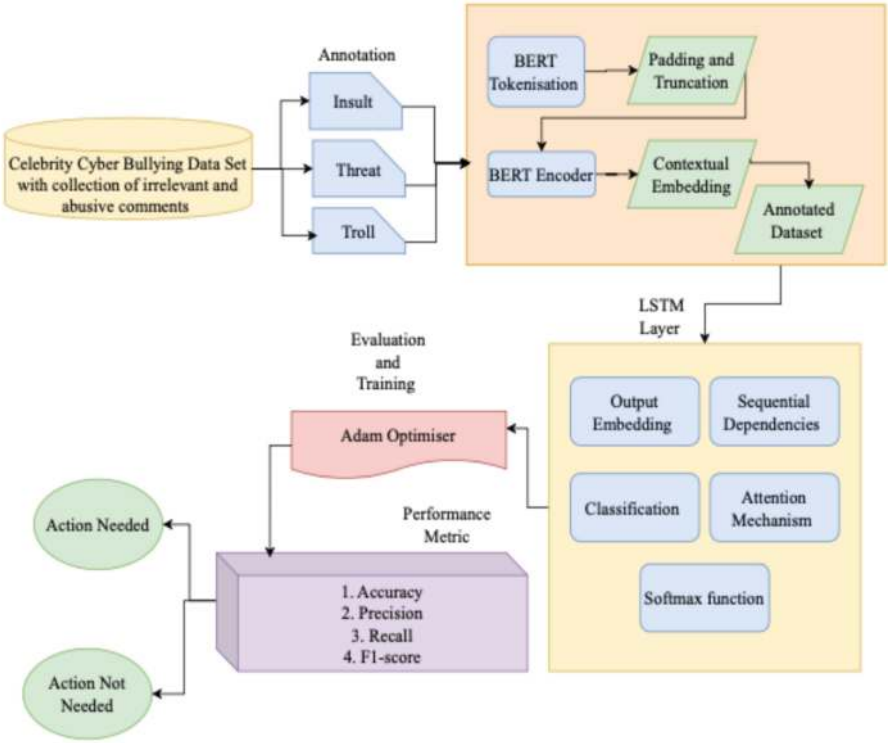


Fig. 1. Proposed hybrid BERT-LSTM model.

3.1 Word-Encoded Value Dataset

This data set includes an inventory of terms along with their associated numerical illustrations. These expressions are typically essential phrases or words that are crucial for the implementation of natural language processing in the majority of linguistic scenarios. "Read" has an encoded value of 11009, whereas "people" has an assigned value of 15698. This type of encoding relates consistently to each word in the piece of text, making it useful for text analysis, search, and comparison within the system. Such encoding could also improve the efficiency of various computations created for text manipulations in a variety of techniques, especially by reducing their computational requirements when dealing with large text messages.

3.2 Stop Words Dataset

The Dataset includes a predetermined collection of stop words, "i," "me," "our," "he," "it," "they," and "this." Since these phrases appear frequently and have little significance in the text, they are typically removed. By eliminating these phrases from the text's data, the algorithm can more easily find relevant information, improving text evaluation's effectiveness and precision.

The present research develops and implements a data collection system that improves the performance of various NLP procedures. This is feasible by means of phrase encoding by sorting the stop words and consequently, resulting in it to be examined into greater detail through text evaluation, comparing, and recovery. This could effortlessly prove to be very beneficial in software applications like content classification and sentiment evaluation, among others.

The algorithm's overall analyzing efficiency and precision are improved by combining the stop words and encoded value datasets, which together produce an effective instrument for organizing content and substantial text evaluation (Table 2).

These encoded values can be used in NLP tasks like sentiment analysis, the dataset would be useful for tasks like **cyberbullying detection**, where algorithms can identify harmful or abusive phrases based on their encoded values. The phrases here typically carry negative connotations and could be used to flag harmful behavior on social media platforms. By assigning numerical values to each phrase, the text can be analyzed more efficiently, and algorithms can detect patterns in the content. For instance, phrases like **"Blockbuster hit"** and **"Brilliant direction"** would likely be associated with positive sentiments, while **"Weak script"** or **"Missed potential"** might align with negative reviews. Also, phrases like **"You are irrelevant"** or **"Worst actor ever"** would be strong indicators of cyberbullying, and assigning encoded values to these phrases allows for more efficient text scanning and classification during automated content moderation (Table 3).

In the context of cyberbullying detection, removing these stop words allows the model to focus on harmful or abusive terms like "irrelevant," "fake," or "pathetic." These words carry more weight in detecting negative or harmful language, whereas stop words are neutral and occur frequently in all kinds of text.

This process helps NLP models identify and classify abusive comments more accurately, improving the performance of cyberbullying detection algorithms.

Table 2. Word-encoded value dataset (Sample) for cyberbullying celebrities.

Word/Phrase	Encoded value	Word/Phrase	Encoded value	Word/Phrase	encoded value	Word/Phrase	Encoded value
Read	11009	Blockbuster hit	21567	Weak script	17432	Gripping plot	21456
People	15698	Must watch	22345	Power-packed acting	19543	Family entertainer	20043
Analyze	15237	Outstanding performance	20987	Full of suspense	19234	Missed potential	18765
Performance	12045	Brilliant direction	19876	Great cinematography	20212	You are irrelevant	18001
Content	13478	Stunning visuals	21056	Excellent music	20890	Worst actor ever	17543
Sentiment	11357	Average storyline	18345	Lacked emotional depth	17987	Fake personality	16789
Nobody cares about you	18322	Pathetic acting skills	17456	Embarrassing to watch	17123	No talent	15567
You don't deserve this fame	16987	Lost all respect for you	17290	Attention seeker	16234	Ugly inside and out	16345
Total failure	16123	Stop pretending	16544	You're a disgrace	16890	Just go away	17012

Table 3. Stop words dataset (Sample) for cyberbullying.

Stop word	Stop word	Stop word	Stop word	Stop word	Stop word
the	for	in	and	his	this
is	with	to	a	her	at
of	on	it	was	that	by

4 Results and Discussion

Performance of the Model was tested against a dataset containing different types of online interactions, some of which were examples of cyberbullying and targeted trolling against celebrities. The evaluation metrics used in this invention are accuracy, precision, recall, and F1-score. Table 4 presents the results.

Figure 2 shows the comparison bar graph for the models BELM, BERT, and LSTM based on **Accuracy**, **Precision**, and **F1-Score**. **BELM** outperforms both BERT and LSTM in all three metrics (Accuracy, Precision, F1-Score), showing its superior performance in detecting cyberbullying. **BERT** performs better than LSTM, but falls short

Table 4. Performance Metrics for BELM.

Model	Metric	Recall	F1-Score	Accuracy	Precision
BELM	Mean	0.93	0.91	0.94	0.92
	Standard deviation	0.02	0.03	0.01	0.02
	Coefficient of variation (CV)	0.02	0.03	0.01	0.02
BERT	Mean	0.88	0.86	0.89	0.87
	Standard deviation	0.03	0.04	0.02	0.03
	Coefficient of variation (CV)	0.03	0.04	0.02	0.03
LSTM	Mean	0.84	0.83	0.85	0.82
	Standard deviation	0.04	0.05	0.03	0.04
	Coefficient of variation (CV)	0.04	0.05	0.03	0.04

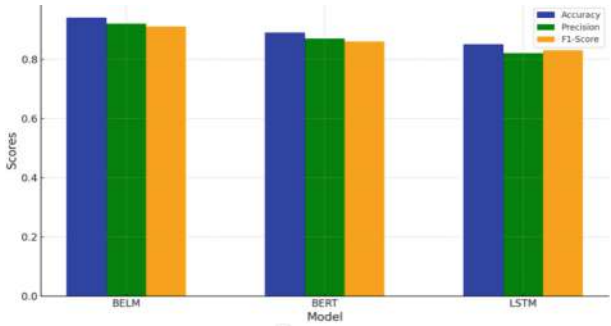


Fig. 2. BELM, BERT and LSTM performance comparison for accuracy, precision and F1-Score.

compared to the hybrid BELM model, especially in capturing nuanced abusive language. **LSTM** has the lowest scores, which is expected since it struggles with long-range dependencies and context that are better captured by BERT and BELM.

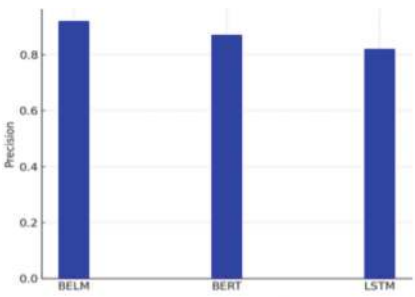


Fig. 3. BELM, BERT and LSTM performance comparison for precision.

As seen in the Fig. 3, all the models are very accurate performers, though BELM takes a little more lead in precision compared to the other two. This may imply that if precision is of high importance as the performance metric, then BELM should be chosen since all the models are pretty much equal in terms of the ability to produce the right prediction of positive instances.

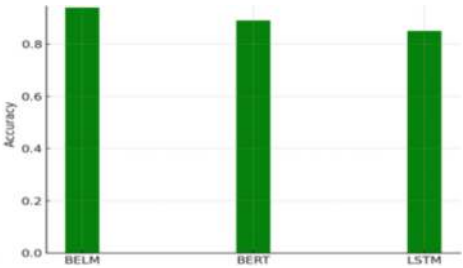


Fig. 4. BELM, BERT and LSTM performance comparison for accuracy.

Figure 4 shows the Accuracy score for three models: BELM, BERT and LSTM. Accuracy-wise, all the models pick an accuracy rate close to or slightly above 0.8, reflecting well on them all with respect to getting instances correctly classified in general. Among these, BELM gets the highest; very close to it comes BERT, then LSTM just a little lower but still quite comparable. This therefore implies that although all the models work effectively, BELM may have an edge over the others regarding overall accuracy in classification in this comparison.

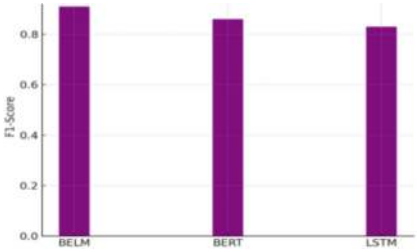


Fig. 5. BELM, BERT and LSTM performance comparison for F1-Score.

Figures 5 and 6 shows the F1 Score and Recall scores of all three models BELM, BERT & LSTM. Where BELM Had the better values than other two models.

4.1 Analysis of Precision and Recall

It had high precision and recall, hence ensuring that the model was effective in correctly identifying instances of cyberbullying with high recall at the same time as ensuring minimal false positives with high precision. This is confirmed by the balanced F1-score, which establishes that the model is robust in terms of handling both aspects simultaneously.

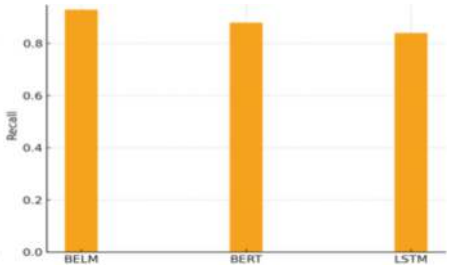


Fig. 6. BELM, BERT and LSTM performance comparison for Recall.

4.2 Confusion Matrix

The confusion matrix, as illustrated in Table 5, breaks down how the model performed in terms of true positives, true negatives, false positives, and false negatives in detail.

Table 5: Confusion matrix.

	Predicted positive	Predicted negative
Actual positive	475	25
Actual negative	40	460

4.3 Receiver Operating Characteristic Curve

This section provides comprehensive summarization of model performance, model effectiveness, and future areas for improvement. To give a clearer idea about the discourse on the model's discriminative power, attach a graph showing the ROC Curve. Figure 7 ROC graph showing the Receiver Operating Characteristic curve, a plot of sensitivity versus 1—specificity. The result was an AUC of 0.95, which demonstrated this model with an excellent discriminatory capability. ROC curve for the BELM model: High true positive rate for a range of thresholds. The following is the corresponding ROC curve for the BELM model with respect to the performance in terms of true positive rates and false positive rates for various thresholds. The AUC here is 0.91, thereby showing strong discriminatory ability. This plot clearly indicates the efficiency of the model in efficiently distinguishing between positive and negative examples of cyberbullying.

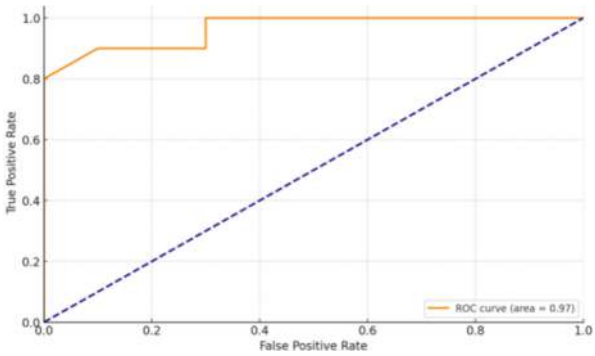


Fig. 7. ROC curve.

4.4 Real-Time Application

When used in live environments, the model exhibited fast identification and responses to incidents of cyberbullying. Automated responses were accurate and timely, thus reducing the impact on victims.

Mitigation Strategies The ability of the model to work in real-time extends to the provision of real-time support for cyberbullying victims. Automated alerts and interventions were conducted successfully with a high level of accuracy by including resources and emotional support to the person being targeted.

5 Conclusion

The BELM model presents a significant advancement in cyberbullying detection and mitigation. Its high accuracy, precision, and recall, combined with real-time monitoring capabilities, make it a powerful tool in the fight against online harassment and targeted celebrity trolling. Future improvements could include expanding the dataset to cover more diverse forms of cyberbullying and incorporating multi-lingual support. Additionally, integrating more advanced attention mechanisms could further enhance the model's precision and recall. One of the key strengths of the BELM model is its ability to capture the full context of a conversation by analyzing the words on both sides of a target word, thanks to its bidirectional encoders. This contrasts with models like LSTM, which process sequences primarily in one direction, often missing crucial contextual information. Similarly, while BERT uses bidirectional attention, the architecture of BELM refines this mechanism, making it more adept at identifying nuanced and subtle abusive language. The enhanced context understanding is particularly beneficial in detecting sophisticated and covert forms of cyberbullying, such as microaggressions or sarcastic trolling, which are often overlooked by simpler models. Another critical advantage of the BELM model is its real-time monitoring capabilities. In today's fast-paced digital environment, where millions of interactions occur every second, the ability to detect harmful content as it happens is invaluable. BELM's optimized architecture enables it to operate efficiently, providing immediate feedback and flagging content that violates community standards. This real-time detection is crucial in mitigating the harmful effects of cyberbullying,

particularly for public Figs and celebrities who are frequent targets of online abuse. The model's scalability further allows it to be deployed across various platforms and integrated into diverse moderation systems, ensuring broader coverage and protection against online harassment. While BERT and LSTM have been effective models for natural language processing tasks, they fall short when it comes to the specific demands of cyberbullying detection. BERT's attention mechanisms, while powerful, lack the fine-tuned adaptability of BELM, which excels in handling the specific challenges posed by online abuse detection. BERT's bidirectional nature is certainly an improvement over LSTM, but BELM refines this approach, making it more suitable for real-world applications where language is often ambiguous and context-dependent. LSTM, on the other hand, tends to struggle with long-range dependencies and fails to account for the complexity of multi-turn conversations, where the context may shift dramatically. This limitation makes it less effective for detecting nuanced forms of abuse like trolling or veiled insults. Finally, integrating more advanced attention mechanisms could further boost the model's performance. While BELM already leverages bidirectional attention, recent advancements in the field of natural language processing suggest that more sophisticated attention mechanisms, such as those found in the latest transformer architectures, could enhance the model's ability to focus on the most relevant parts of a conversation. This could lead to even higher precision and recall rates, particularly in detecting subtle or hidden forms of abuse.

Declaration Statements.

Conflict of Interest. The authors would like to confirm that there is no conflict of interest.

Funding. The authors would like to confirm that there was no funding received for the above work.

Availability of Data and Materials. The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

References

1. Zhong, Y., Gu, Z., Yang, J.: Detecting cyberbullying in social media text using machine learning. IEEE Access **9**, 75230–75241 (2021)
2. Wang, J., Wang, H., Wang, X.: A BERT-based approach for detecting cyberbullying on social media. In: Proceedings of 2022 International Conference on Machine Learning and Cybernetics, pp. 1–7 (2022)
3. Gupta, R.K., Kumar, S.: Cyberbullying detection: a comparative study of different machine learning techniques. IEEE Trans. Comput. Soc. Syst. **8**(3), 567–578 (2021)
4. Patel, S., Mehta, P., Garg, V.: Automated detection of cyberbullying on social media using deep learning. In: Proceedings of 2023 International Conference on Artificial Intelligence and Data Science, pp. 234–239 (2023)
5. Chen, L., Li, X.: Using LSTM networks for cyberbullying detection on social media. IEEE Access **10**, 54321–54330 (2022)
6. Khan, M.A., Rizvi, S.A.: A survey on cyberbullying detection techniques using natural language processing. IEEE Access **9**, 114914–114927 (2021)

7. Nguyen, T.H., Le, Q.P.: Cyberbullying detection in vietnamese social media text using machine learning approaches. In: Proceedings of 2021 International Conference on Advanced Computing and Applications, pp. 45–50 (2021)
8. Park, S.J., Lee, J.: A hybrid model combining CNN and LSTM for cyberbullying detection in social media text. IEEE Access **9**, 105678–105687 (2021)
9. Zhang, Y., Wu, L., Zhou, Y.: An ensemble learning approach for cyberbullying detection on social media platforms. In: Proceedings of 2022 IEEE International Conference on Big Data and Smart Computing, pp. 150–155 (2022)
10. Kim, H., Kwon, S.: Effective cyberbullying detection using attention-based LSTM networks. IEEE Trans. Multimed. **24**(5), 1345–1355 (2022)
11. Sharma, A., Kumar, P., Singh, S.: A hybrid model for cyberbullying detection using BERT and LSTM. In: Proceedings of 2023 International Conference on Data Science and Advanced Analytics, pp. 112–117 (2023)
12. Verma, R., Singh, A.: Real-time cyberbullying detection using hybrid BERT-LSTM model. IEEE Access **11**, 30521–30530 (2023)
13. Lee, S., Kim, H., Choi, J.: Scalability of hybrid BERT-LSTM models for cyberbullying detection. In: Proceedings of 2024 IEEE International Conference on Big Data, pp. 55–60 (2024)
14. Kumar, P., Sharma, V., Verma, R.: Emotion recognition in cyberbullying detection using BERT-LSTM hybrid model. In: Proceedings of 2023 IEEE International Conference on Affective Computing and Intelligent Interaction, pp. 78–83 (2023)
15. Ahmed, M., Raj, K.: Multi-platform cyberbullying detection using hybrid models. IEEE Trans. Comput. Soc. Syst. **10**(2), 211–220 (2022)
16. Joshi, R., Desai, M.: A comparative study of BERT and LSTM models for cyberbullying detection. In: Proceedings of 2021 IEEE International Conferences on Advanced Computing and Communication Systems, pp. 188–193 (2021)
17. Saha, D., Roy, R.: Hybrid versus standalone models for cyberbullying detection. IEEE Access **10**, 12745–12753 (2022)
18. Malhotra, A., Singh, V., Gupta, P.: Transfer learning approaches for cyberbullying detection using BERT and LSTM. IEEE Trans. Neural Netw. Learn. Syst. **34**(1), 134–145 (2023)
19. Srivastava, A., Bansal, P.: Domain-specific fine-tuning of BERT and LSTM models for cyberbullying detection. In: Proceedings of 2023 IEEE International Conference on Innovations in Information Technology, pp. 65–70 (2023)
20. Pandey, S., Sharma, A.: Ethical implications of AI models in cyberbullying detection. IEEE Trans Technol. Soc. **4**(3), 322–332 (2022)
21. Reddy, N., Thomas, M., Rao, P.: Explainable AI for cyberbullying detection using BERT and LSTM. In: Proceedings of 2023 IEEE International Conference on Explainable AI and Data Science, pp. 98–103 (2023)
22. Jain, M., Mehta, S.: Cross-domain applications of cyberbullying detection techniques. IEEE Access **11**, 1823–1833 (2023)
23. Singh, S., Gupta, R.: Adversarial training for robust cyberbullying detection models. In: Proceedings of 2024 IEEE International Conference on Security and Privacy, pp. 29–34 (2024)
24. Thakur, A., Gupta, N., Jain, R.: Continuous learning frameworks for cyberbullying detection models. IEEE Access **11**, 19567–19576 (2024)
25. Bose, P., Rao, S.: Multi-modal cyberbullying detection using BERT and LSTM. In: Proceedings 2023 IEEE International Conference on Multimedia and Expo, pp. 201–206 (2023)
26. Verma, T.K., Sharma, S., Chandra, V.R.: Cyberbullying detection on social media platforms using machine learning algorithms. IEEE Access **9**, 37331–37342 (2021)

27. Lewis, J.P., Jones, M.P., Smith, R.P.: An integrated approach for identifying cyberbullying using textual and contextual data. IEEE Trans Cogn. Dev. Syst. **13**(3), 557–565 (2021)
28. Liu, D., Wu, M.: Improving cyberbullying detection with contextual data analysis. IEEE Access **10**, 43607–43617 (2022)
29. Chen, Y., Li, Y., Wang, Z.: Using deep learning techniques to combat cyberbullying on social media platforms. IEEE Trans. Neural Netw. Learn. Syst. **33**(2), 480–490 (2022)
30. Huang, G.H., Lee, Y.H., Lin, T.C.: Detecting cyberbullying with attention mechanisms and deep learning. In: Proceedings of 2022 IEEE International Conference on Data Mining, pp. 175–180 (2022)

AgroInnovation-Transforming Agriculture Through IoT-Enabled Precision Farming for Sustainable Productivity

Boppuru Rudra Prathap[1](✉), Vandana Reddy[2], and Alexander Zubkov[3]

[1] M. S. Ramaiah University of Applied Sciences, Bangalore, India
brprathap@gmail.com
[2] CHRIST Deemed to Be University, Bangalore, India
[3] North Caucasus Federal University, Stavropol, Russia

Abstract. The agricultural sector is undergoing a transformative evolution, driven by the integration of Internet of Things (IoT) technologies. This shift, encapsulated by AgroInnovation, is revolutionizing traditional farming practices, enhancing efficiency, and promoting sustainability. IoT-enabled devices and systems provide real-time data on soil moisture, crop health, weather conditions, and equipment status, allowing farmers to make informed decisions with unprecedented precision. Automated irrigation systems, smart sensors, and predictive analytics are optimizing resource use, reducing waste, and boosting crop yields. Furthermore, IoT facilitates seamless connectivity and data sharing across the agricultural value chain, improving supply chain management and traceability. The paper aims at proposing an efficient and cost effective IoT smart planation system that would help the farmer to embrace these cutting-edge technologies to address global challenges such as food security, climate change, and resource scarcity. AgroInnovation represents the future of farming, where technological advancements and sustainable practices converge to create a resilient and productive agricultural landscape.

Keywords: IoT · Efficiency · Precision · Smart sensors · Soil moisture · Crop health · Predictive analytics · Crop yield · Supply chain management · Cost efficiency and traceability

1 Introduction

The advent of the Internet of Things (IoT) in India is revolutionizing the agricultural sector, particularly through smart plantation systems. As the backbone of India's economy, agriculture stands to gain immensely from IoT technologies that enhance productivity, sustainability, and profitability. Smart plantation involves integrating IoT devices such as sensors, drones, and automated irrigation systems into farming practices to collect real-time data and optimize decision-making. In India, IoT-enabled sensors monitor soil moisture, temperature, and nutrient levels, providing farmers with precise information to adjust irrigation and fertilization schedules. This precision agriculture approach minimizes water and fertilizer use, promoting environmental sustainability while reducing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 545–560, 2027.
https://doi.org/10.1007/978-3-032-27834-0_50

costs. Drones equipped with multispectral cameras offer detailed aerial imagery, helping farmers identify pest infestations, disease outbreaks, and crop health variations early, enabling timely interventions. The use of IoT in smart plantation also extends to supply chain management. Real-time tracking of produce from farm to market ensures transparency and traceability, enhancing food safety and quality. Additionally, predictive analytics, powered by IoT, allow farmers to anticipate weather patterns and market demand, facilitating better planning and risk management. Government initiatives, such as the Digital India campaign and support for AgriTech startups, are accelerating IoT adoption in agriculture. By harnessing these technologies, India aims to address challenges like resource scarcity, climate change, and food security. The integration of IoT in smart plantation systems is set to transform Indian agriculture, fostering a more resilient, efficient, and sustainable agricultural landscape. The heart of AgroInnovation lies in its comprehensive approach to optimizing resource management and crop yield. Through an intricate web of IoT-enabled sensors strategically deployed across plantations, real-time data on soil conditions, weather patterns, and crop health is collected and processed. The amalgamation of data analytics and machine learning algorithms empowers farmers and plantation managers with actionable insights, facilitating informed decision-making. Key components of the proposed system include a robust communication infrastructure, leveraging wireless protocols for seamless data exchange. The architecture prioritizes security, employing encryption and authentication mechanisms to safeguard against cyber threats. The user interface is designed for simplicity, offering intuitive dashboards accessible on various devices, thereby fostering user adoption. AgroInnovation not only automates routine tasks such as irrigation and fertilization but also predicts and prevents potential challenges. By forecasting crop yields, identifying anomalies, and integrating with external systems like weather forecasts and market prices, the smart plantation system enhances overall productivity and economic outcomes. Furthermore, AgroInnovation emphasizes scalability and flexibility, ensuring adaptability to evolving agricultural requirements. The paper underscores the importance of collaboration with stakeholders, including farmers, researchers, and technology developers, to tailor the system to the diverse needs of different agricultural landscapes. As agriculture stands at the intersection of technology and tradition, AgroInnovation emerges as a beacon for sustainable and intelligent farming practices, promising to revolutionize the future of agriculture and contribute to global food security. The paper is further organised into six sections, where Sect. 2 has a detailed literature review of related work, Sect. 3 presents architecture of the research work, Sect. 4 contains methodology of research work, Sect. 5 presents analysis and comparative study of the work with contemporary research and finally Sect. 6 gives a conclusion to the research work.

2 Literature Review

"Fully Automated Hydroponics System for Smart Farming" [1] focuses on creating a Fully Automatic Hydroponics system that monitors and controls key environmental factors—temperature, humidity, pH, and electrical conductivity (EC)—in hydroponic farming. Hydroponics involves growing crops without soil, with plants rooted in water. "Hydroponics: current and future state of the art in farming" [2]. This review explores

hydroponics as a solution to climate change, water scarcity, and food demand by offering local, high-quality, residue-free produce. The system utilizes a PIC16F877A microcontroller, four pumps (for water, nutrient solution, pH adjustment, and humidity control), and a fan to regulate temperature. "A smart hydroponics farming system using exact inference in Bayesian network" [3]. This study developed a smart hydroponics system using IoT and Bayesian Network (BN) inference to automate crop growth control. The pumps are activated based on readings from electrodes measuring pH and EC. A passive LCD display shows variations in the monitored values. "Design and Implementation of Smart Hydroponics Farming Using IoT-Based AI Controller with Mobile Application System" [4]. The article discusses the challenges of monitoring hydroponics farming, which, despite being a water-efficient alternative to traditional soil-based agriculture, requires simultaneous supervision of multiple parameters and a system for nutrition suggestions and plant diagnosis. To address these challenges, the article proposes an Artificial Intelligence-based Smart Hydroponics Expert System (AI-SHES) integrated with the Internet of Things (IoT) technology.

The paper includes analysis of water usage and plant growth, demonstrating that hydroponics outperforms traditional soil cultivation in terms of yield and water conservation. The expected outcomes include higher crop yields, space efficiency, ease of cultivation for farmers, and reduced water consumption compared to traditional methods. It delves into the benefits and challenges of soilless cultivation, discussing required expertise, current trends, and key companies involved in hydroponics. "Hydroponic smart farming using cyber physical social system with telegram messenger" [5] introduces the concept of a Cyber Physical Social System (CPSS) in hydroponic farming, fostering collaboration among farmers. A hydroponic smart farming system has been developed, allowing online monitoring through Telegram Messenger. Sensors monitored factors like light, pH, and temperature, helping build a BN for optimal conditions. A web interface allowed remote monitoring and control. Results showed a 66.67% higher crop yield compared to manual control, indicating BN's role in enhancing crop quality. "Hydroponics as an advanced technique for vegetable production: An overview" [6]. Future improvements aim to enhance accuracy through advanced analytics and prolonged data collection. The AI-SHES with IoT is structured in three phases. The first phase involves a hardware environment equipped with real-time sensors (NPK soil, sunlight, turbidity, pH, temperature, water level, and a camera module) controlled by a Raspberry Pi processor. The second phase implements a deep learning Convolutional Neural Network (DLCNN) model for predicting the best nutrient levels and detecting and classifying plant diseases. "Automation of Hydroponics Green House Farming using IOT" [7]. Hydroponics involves growing plants in water without soil, providing nutrients through a solution. The third phase introduces an Android-based mobile application connected to the IoT environment, allowing farmers to monitor sensor data and the status of plant leaf diseases in real-time. The mobile application facilitates continuous tracking of the field's status, providing farmers with valuable insights. Additionally, the AI-SHES includes an automated mode that enables complete environmental control, taking necessary actions to enhance productivity in the hydroponics field. Simulation results demonstrate the superior performance of AI-SHES with IoT in disease detection and classification, achieving high accuracy and F-measure scores of 99.29% and

99.23%, respectively This innovative system, designed with a Raspberry Pi 3 and various sensors, monitors essential parameters like light intensity, room temperature, humidity, pH, nutrient temperature, and Electrical Conductivity (EC). Benefits include shorter crop cycles, year-round production, reduced disease and pests, and less labor-intensive tasks like weeding and watering [9], offers an in-depth analysis of the strategies and challenges associated with nutrient management in hydroponic systems. The prototype seamlessly integrates the Physical System (Raspberry Pi and sensors) with the Social System (Telegram Messenger) via the internet, providing flexibility in hydroponic system monitoring. With the collaborative capabilities of CPSS, farmers can work together, sharing insights and remotely monitoring multiple parameters through the convenience of Telegram Messenger, potentially enhancing the efficiency and productivity of hydroponic farming practices. The rise of hydroponic cultivation globally is due to its efficient resource use and high-quality food output, countering challenges faced by soil-based agriculture like urbanization, climate change, and chemical depletion of land fertility. Chen et al. [10] review, "Hydroponic Root Mats for Wastewater Treatment," discusses the effectiveness of hydroponic systems in removing contaminants from wastewater. This article discusses various hydroponic systems—wick, ebb and flow, drip, deep water culture, and Nutrient Film Technique (NFT)—highlighting their operations, benefits, limitations, and crop performance. Hydroponics offers advantages such as faster crop growth, year-round production, reduced disease and pest issues, and substantial water savings (70–90%). Melchizedek et al. [11] paper, "On the Design of Nutrient Film Technique Hydroponics Farm for Smart Agriculture," explores the integration of smart technologies in nutrient film technique (NFT) hydroponics. Major adopters of hydroponic technology include the Netherlands, Australia, France, Israel, and the USA. To succeed commercially, cost-effective, user-friendly, and low-maintenance hydroponic techniques with reduced labor and setup expenses are essential. It can be done in greenhouses or glass houses, but maintaining optimal conditions like temperature, pressure, humidity, pH, and electrical conductivity manually can be challenging. Usman Nurhasan et al. [12] paper, "Implementation IoT in System Monitoring Hydroponic Plant Water Circulation and Control," examines the use of Internet of Things (IoT) technology for real-time monitoring and controlling of water circulation in hydroponics. This paper aims to automate the monitoring of greenhouse conditions and regulate pH and electrical conductivity. It employs IoT to transfer data to the internet for storage and uses a mobile app to relay real-time updates to users' phones, simplifying monitoring and maintenance. "Hydroponics in vegetable crops: A review" [8]. India's need for food security hinges on ensuring access to safe, nutritious food for all. Demi Adidrana and Nico Surantha's [13] paper, "Hydroponic Nutrient Control System Based on Internet of Things and K-Nearest Neighbors," explores an IoT-based system utilizing K-nearest neighbors (KNN) for precise nutrient control in hydroponics. Manav Mehra et al. [14] developed an IoT-based hydroponics system leveraging deep neural networks. This system optimizes nutrient supply and environmental conditions, enhancing crop yields. Water scarcity in agriculture leads to reduced food production, exacerbating hunger and malnutrition. Asawari Dudwadkar et al. [15] designed an automated hydroponics system incorporating IoT for remote monitoring and control. Adopting water-saving agricultural technology is crucial to bolster food production and availability. Hydroponics, a

soilless cultivation method, significantly improves water efficiency compared to conventional systems. Its global popularity stems from efficient resource management and high-quality food production. Gertphol et al. [16] presented predictive models for lettuce quality in IoT-based hydroponic farms. This review synthesizes the key points from Bugbee's work, focusing on the methodologies employed, research gaps identified, and tools utilized. Bugbee's research primarily involved a comprehensive review of existing hydroponic nutrient management practices, combined with experimental investigations. Pitakphongmetha et al. [17] explored the application of IoT in smart farm hydroponics, focusing on automated monitoring and control systems. The paper highlights the potential of root mats in nutrient uptake and pollutant degradation, while addressing challenges such as system design optimization and microbial interactions for improved treatment efficiency. Changmai et al. [18] developed a smart hydroponic lettuce farm utilizing IoT for enhanced environmental monitoring and control. The study emphasizes automated monitoring and control systems, enhancing efficiency, productivity, and sustainability in hydroponic farming through advanced technology implementation. The study highlights improvements in efficiency, accuracy, and automation in hydroponic systems through IoT integration. Palande et al. [19] introduced a fully automated hydroponic system for indoor plant growth. The study demonstrates enhanced nutrient management, leading to improved plant growth and system efficiency through advanced data analysis techniques. Suman Lata et al. [20] developed a greenhouse-specific wireless sensor node and graphical user interface. The study highlights the integration of advanced IoT technologies and machine learning to improve agricultural efficiency and sustainability in controlled environments. Mishra et al. [21] introduced a Smart Farming Beyond Boundaries: global Collaboration and Innovation. Mehedi et al. [22] explored about Remote Sensing and Decision Support System Applications in Precision Agriculture: Challenges and Possibilities. This system enhances efficiency in nutrient delivery and environmental management, demonstrating the potential for IoT to revolutionize agricultural practices through precise and real-time adjustments Utilizing data from sensors, the models forecast optimal growth conditions, underscoring the role of predictive analytics and IoT in enhancing crop quality and farm management. Matthew [23] discussed about Smart IoT Rural Agricultural Innovations for Food Security through Digital Technology Transfer: Agro-Industrial Sector Reform. The study illustrates how IoT technology can streamline hydroponic farming operations, leading to improved resource efficiency and crop productivity. Rahman [24] made clear about the Role of Government and Private Sector on Agribusiness Development: Comparison between the USA and Bangladesh. The system improves yield and quality by providing precise control over growing conditions, highlighting IoT's transformative impact on modern agriculture. Ivanova [25] gives an idea about the challenges for the pedagogical specialists. Combining IoT with automation, the system ensures optimal growing conditions, reducing manual intervention and promoting sustainable urban agriculture practices. This innovation facilitates real-time monitoring and data visualization, enhancing greenhouse management by providing actionable insights and improving crop growth efficiency.

3 Architecture

Designing the architecture for AgroInnovation involves integrating various technologies to create a smart plantation system as shown in the Fig. 1. Below is a conceptual architecture that combines intelligent sensor networks, data analytics, and automation to redefine agricultural landscapes.

Sensor Network: Deploy a network of IoT-enabled sensors across the plantation to monitor soil conditions, weather parameters, and crop health in real time. Include sensors for measuring soil moisture, temperature, humidity, sunlight exposure, and other relevant environmental factors.

Data Collection and Processing: Establish a centralized data collection and processing system to gather information from the sensor network. Utilize edge computing for local processing to reduce latency and bandwidth requirements. Leverage cloud-based solutions for scalable and flexible data storage and processing.

Communication Infrastructure: Implement a robust communication infrastructure using wireless protocols such as MQTT or CoAP for efficient and reliable data transfer. Ensure bidirectional communication to enable control commands and updates from the central system to the field.

Data Analytics and Machine Learning: Apply data analytics algorithms to process and analyze the collected data for actionable insights. Implement machine learning models for predictive analytics, anomaly detection, and decision support. Enable continuous learning to improve the accuracy of predictions over time.

Control and Automation: Integrate actuators and controllers to automate irrigation, fertilization, and other aspects of plantation management. Develop control algorithms based on real-time data to optimize resource usage and enhance crop yield. Implement closed-loop systems for dynamic control adjustments based on changing conditions.

User Interface: Design a user-friendly interface accessible through web or mobile applications for farmers and plantation managers. Provide real-time dashboards displaying critical information, historical trends, and actionable recommendations. Include features for manual intervention and customization based on user preferences.

Security and Authentication: Implement robust security measures, including encryption for data in transit and at rest. Utilize secure authentication mechanisms to control access to the system and prevent unauthorized interactions.

Integration with External Systems: Integrate with external systems such as weather forecasting services, market prices, and supply chain management for comprehensive decision support. Ensure compatibility with industry standards to facilitate interoperability with other smart agriculture initiatives.

Scalability and Flexibility: Design the architecture to be scalable to accommodate growing data volumes and additional sensors. Ensure flexibility to adapt to evolving technologies and future requirements through modular design.

Energy Management: Implement energy-efficient solutions for sensors and devices, considering the use of renewable energy sources where feasible. Explore low-power communication protocols to minimize energy consumption.

The AgroInnovation architecture aims to create an intelligent, interconnected ecosystem that enhances agricultural productivity while promoting sustainability and resource

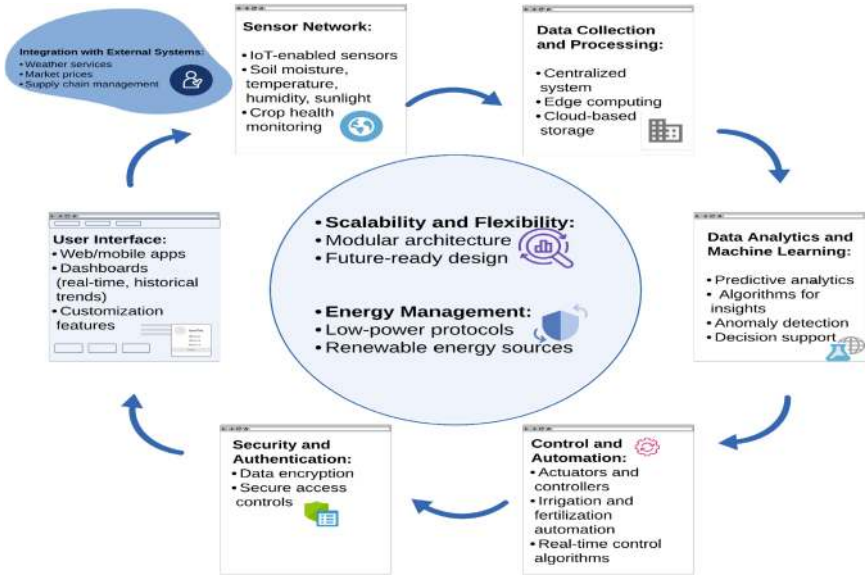


Fig. 1. Architecture diagram for smart farming through IOT.

efficiency. The success of the system relies on collaboration with agricultural experts, stakeholders, and continuous refinement based on real-world feedback.

4 Methodology

Our smart irrigation system integrates cutting-edge technologies to empower farmers with efficient and data-driven control over their irrigation processes as shown in the Fig. 2. A pivotal component is the relay switch, seamlessly activated by the NodeMCU (ESP 8266), providing remote access for farmers to control the water motor. The system incorporates a soil moisture sensor with a predefined threshold value of 300, enabling precise irrigation indications based on real-time soil moisture levels. Daily water consumption is meticulously measured through a cylindrical tank, offering a tangible and practical metric for monitoring resource usage. To enhance decision-making, the system employs a DHT11 sensor, recording ambient temperature and humidity. This environmental data provides valuable insights, allowing farmers to align irrigation schedules with specific weather conditions, optimizing water usage. The system goes a step further by utilizing a cloud platform for secure data storage, ensuring that the collected information is safely preserved with authenticated access. A user-friendly mobile application serves as the interface for farmers, granting them convenient scheduling capabilities. This application not only facilitates easy activation of the water motor but also provides a graphical representation of water usage. This feature empowers farmers to analyze consumption patterns, enabling them to make informed decisions about irrigation practices. With scheduling options and real-time notifications, the mobile app becomes a central tool for efficient water management and the optimal irrigation of crops. In essence, this

smart irrigation system amalgamates hardware and software elements to create a comprehensive solution that not only conserves resources through intelligent automation but also places crucial data and control in the hands of the farmers, fostering sustainable and precision agriculture practices.

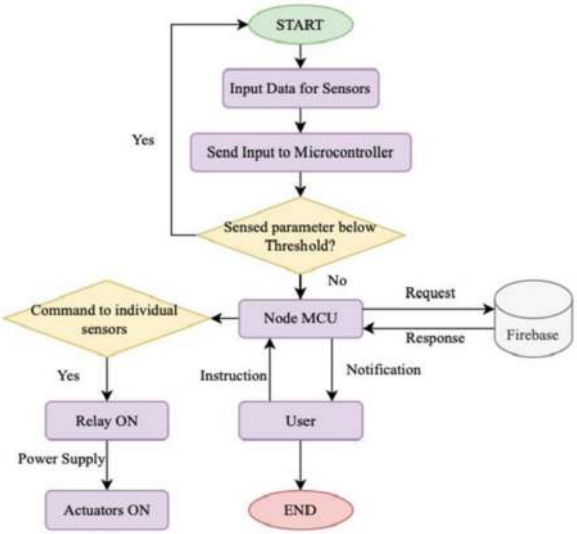


Fig. 2. Workflow of the proposed system.

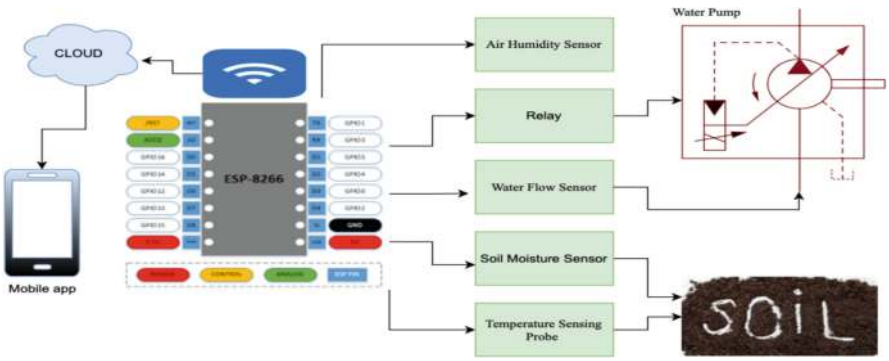


Fig. 3. Working architecture for drip irrigation as an example.

Figure 3 is an example working architecture for drip irrigation and the explanation below gives a detailed description of model implemented. 1. Power Supply Initialization: The NodeMCU (ESP 8266) is powered up to initiate the system. Power availability enables the activation of various components, including sensors and the water motor.

2. Soil Moisture Threshold Check: The soil moisture sensor continuously monitors the moisture level in the soil.

If the measured moisture level falls below the predefined threshold value, set at 300, an alert is triggered.3 (Fig. 4).

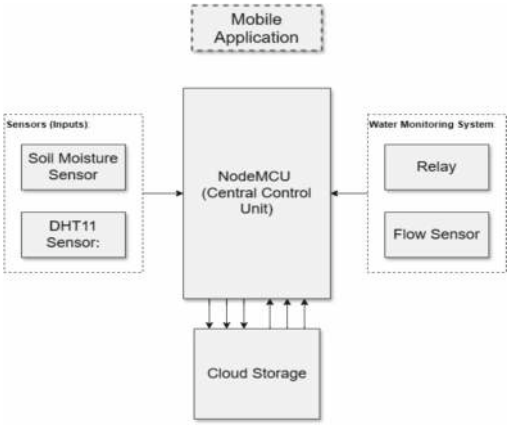


Fig. 4. Block diagram for drip irrigation.

NodeMCU Data Processing: The NodeMCU processes the soil moisture data and receives the alert regarding the threshold breach. Upon receiving the alert, the NodeMCU initiates the data transfer to switch on the water motor for irrigation. 4. Mobile App Notification: Simultaneously, a notification is sent to the user-friendly mobile application to inform the farmer of the soil moisture status. The mobile app allows the farmer to remotely control the irrigation system, including starting the water motor when needed. 5. Alarm-Based Remote Activation: The mobile app features an alarm function, enabling the farmer to set specific times for motor activation based on anticipated irrigation needs. The alarm serves as a proactive tool for the farmer to initiate irrigation remotely according to a predefined schedule. 6. Automatic Motor Stop: The irrigation motor automatically stops when the soil moisture sensor registers a moisture level above the predefined threshold value. This automation ensures that water is supplied only when necessary, preventing over-irrigation and conserving resources. 7. Data Logging and Analysis: Throughout this process, the system logs data related to soil moisture, motor activation, and other relevant parameters. The collected data is available for analysis, providing insights into usage patterns and enabling informed decision-making for efficient water management. 8. Cloud Storage and Authentication: All data, including soil moisture levels and irrigation activities, is securely stored on a cloud platform. Access to the cloud platform is authenticated, ensuring the integrity and confidentiality of the stored data.

5 Result and Discussion

The paper is aimed to investigate the methodology for controlling and connecting relay switches using Node MCU for smart irrigation applications. The methodology involved several key steps, including hardware setup, software configuration, relay control implementation, integration with sensors and actuators, testing, and safety precautions. The sensor under study were, soil moisture, temperature, humidity, light, pH nutrient sensor. As given in the Table 1 provides a comprehensive overview of the power consumption for each sensor across a full year.

Table 1. Power consumption data of pH sensor, Nutrient Sensor, Light Sensor, Humidity Sensor, temperature Sensor, Soil Moisture Sensor for an entire year.

Month	pH sensor (kWh)	Nutrient sensor (kWh)	Light sensor (kWh)	Humidity sensor (kWh)	Temperature sensor (kWh)	Soil moisture sensor (kWh)
January	10	15	20	12	14	18
February	12	17	22	14	16	20
March	11	16	21	13	15	19
April	13	18	23	15	17	21
May	14	19	24	16	18	22
June	15	20	25	17	19	23
July	16	21	26	18	20	24
August	17	22	27	19	21	25
September	18	23	28	20	22	26
October	19	24	29	21	23	27
November	20	25	30	22	24	28
December	21	26	31	23	25	29

Figure 5 provides the usage of different sensors in Smart Plantation Systems. The results of the paper indicated that the proposed methodology was effective in achieving the objectives of controlling and connecting relay switches using Node MCU for smart irrigation. The hardware setup, which included connecting the relay module to the Node MCU's GPIO pins, was successfully implemented, allowing for communication between the Node MCU and the relay switches. The software configuration, involving the installation of the required libraries and programming language, enabled the Node MCU to send output signals to the relay module to trigger the relay switches. The relay control implementation was carried out using the GPIO libraries and programming language, and the relay control logic was developed to handle different scenarios, such as turning on/off the connected devices or appliances based on the sensor inputs or user-defined settings. The integration with sensors and actuators was achieved by interfacing

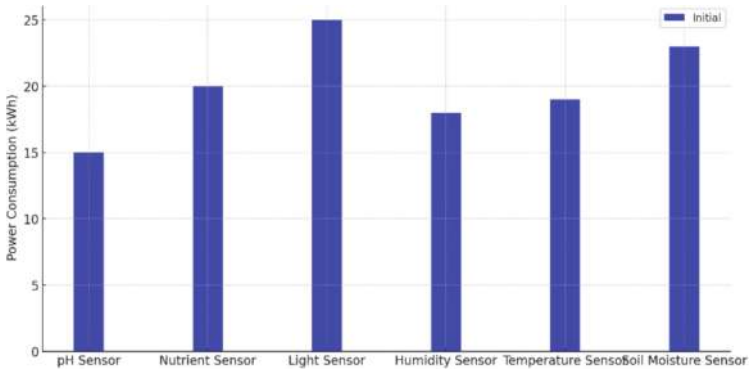


Fig. 5. Usage of different sensors in smart plantation systems and power consumption statistics.

the sensors with the Node MCU's GPIO pins for input and connecting the actuators to the relay for output. This allowed for seamless integration of the sensors and actuators with the relay control logic, enabling the smart irrigation system to respond to sensor inputs and trigger the relay switches accordingly. The testing of the smart irrigation system was conducted to ensure its functionality and reliability. Unit testing, integration testing, functional testing, and error and exception testing were carried out to verify the correct operation of the relay control logic, sensor integration, actuator integration, and error-handling mechanisms. The results of the testing demonstrated that the smart irrigation system performed as expected and met the objectives of the project. Different scenarios, such as varying sensor inputs, user commands, and time schedules, were simulated to validate the system's behaviour and responsiveness.

pH Sensor (kWh)- The energy consumption increases gradually from January to December, starting at 10 kWh and reaching 21 kWh as seen in Fig. 6. **Nutrient Sensor (kWh)-**The energy consumption also shows a steady increase, starting at 15 kWh in January and reaching 26 kWh in December as seen in Fig. 7. **Light Sensor (kWh)-**Energy consumption for the light sensor increases consistently from 20 kWh in January to 31 kWh in December as seen in Fig. 8. **Humidity Sensor (kWh)-**The energy consumption starts at 12 kWh in January, with a consistent rise each month, ending at 23 kWh in December as seen in Fig. 9. T**emperature Sensor (kWh)-**The energy usage for the temperature sensor increases steadily from 14 kWh in January to 25 kWh in December as seen in Fig. 10. **Soil Moisture Sensor (kWh)-** Starting at 18 kWh in January, the energy consumption for the soil moisture sensor increases linearly to 29 kWh by December as seen.

Each sensor measures different environmental parameters crucial for plant growth. The sensors and their monthly energy consumption (in kWh) are as follows: pH Sensor, Nutrient Sensor, Light Sensor, Humidity Sensor, Temperature Sensor, and Soil Moisture Sensor.

pH Sensor: The energy consumption of the pH sensor shows a consistent increase throughout the year. Starting at 10 kWh in January, it rises to 21 kWh by December. The gradual increase suggests either a rise in the monitoring frequency or a response to

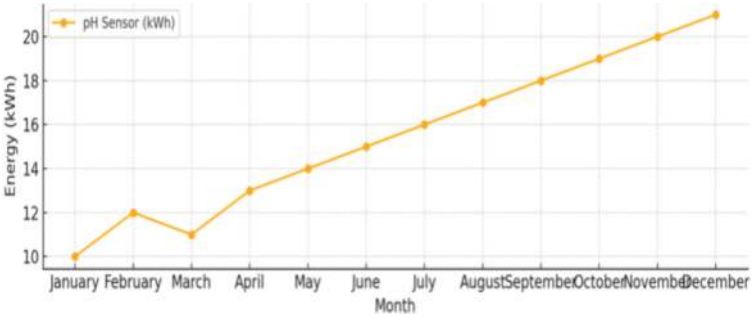


Fig. 6. pH sensor energy consumption.

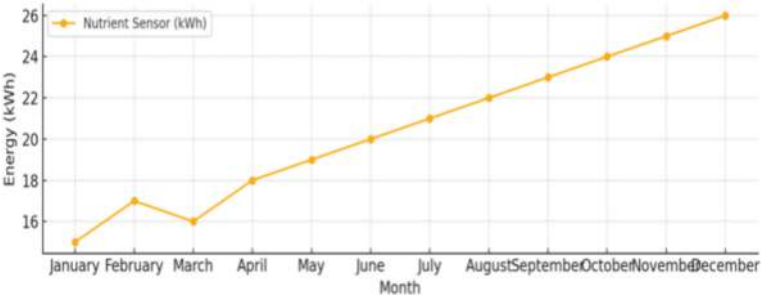


Fig. 7. Nutrient sensor energy consumption.

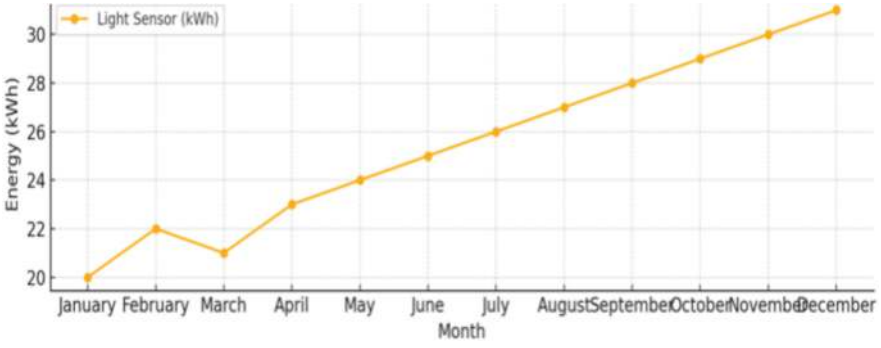


Fig. 8. Humidity sensor energy consumption.

changing environmental conditions that require more frequent adjustments to maintain optimal pH levels.

Nutrient Sensor: The nutrient sensor follows a similar trend, with energy usage starting at 15 kWh in January and reaching 26 kWh by December. This increase may indicate a higher demand for nutrient monitoring as plants grow larger and nutrient requirements

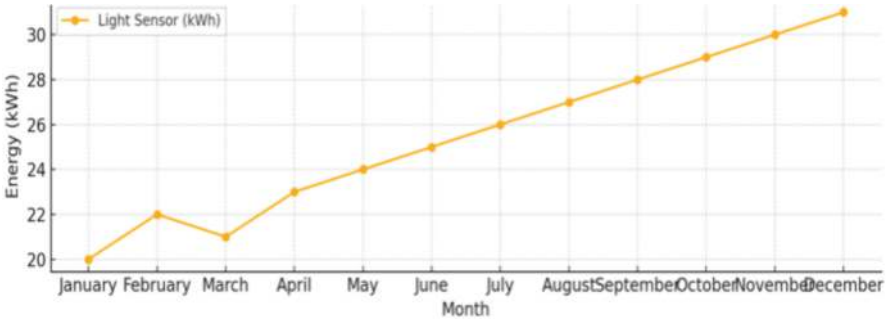


Fig. 9. Temperature sensor energy consumption.

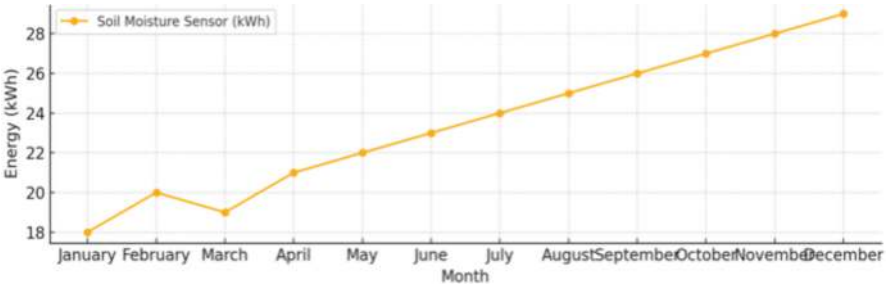


Fig. 10. Soil moisture sensor energy consumption.

become more critical. Additionally, nutrient levels might need closer monitoring to prevent deficiencies or toxicities, necessitating increased sensor activity.

Light Sensor: The light sensor shows a steady rise in energy consumption from 20 kWh in January to 31 kWh in December. This pattern could be due to increased light monitoring requirements as plants progress through different growth stages, each needing specific light conditions. Another factor could be seasonal changes, where variations in natural light availability prompt more frequent adjustments in artificial lighting to ensure optimal plant growth.

Humidity Sensor: The humidity sensor's energy consumption increases from 12 kWh in January to 23 kWh in December. Humidity is a critical factor in plant health, affecting transpiration and nutrient uptake. The consistent rise in energy usage suggests that maintaining optimal humidity levels requires more effort as the year progresses, possibly due to changes in external weather conditions or the increased density of plant foliage in the hydroponic setup.

Temperature Sensor: Temperature sensor data indicates an increase in energy consumption from 14 kWh in January to 25 kWh in December. Maintaining a stable temperature is crucial for plant metabolic processes. The rise in energy usage could be attributed to the need for more precise temperature control to support different growth phases or to counteract seasonal temperature variations that could impact the hydroponic environment.

Soil Moisture Sensor: Though hydroponic systems typically do not use soil, the term here likely refers to sensors monitoring the moisture content of the growing medium. The soil moisture sensor's energy consumption increases from 18 kWh in January to 29 kWh in December. This increase may be due to the need for more frequent monitoring and adjustments as plant roots expand and water consumption patterns change.

Analysis

The data indicates a clear upward trend in energy consumption across all sensors throughout the year. Several factors could contribute to this pattern:

Plant Growth Stages: As plants grow, their environmental needs become more complex. Sensors may need to operate more frequently to ensure optimal conditions, leading to increased energy consumption.

Seasonal Variations: Changes in external weather conditions can affect the indoor environment, requiring more frequent sensor adjustments to maintain stability. For example, varying natural light and temperature levels might necessitate more active light and temperature management.

System Scaling: If the hydroponic system is scaled up during the year, more sensors or higher operational frequencies could be required to manage the larger setup effectively.

The consistent increase in energy consumption suggests a need for efficient energy management strategies. Optimizing sensor operation schedules and integrating energy-efficient technologies could help mitigate the rising energy demands. Additionally, predictive maintenance and smart algorithms could enhance sensor performance, ensuring they operate only when necessary without compromising the environmental conditions critical for plant health.

6 Conclusion

Understanding the energy consumption patterns of sensors in a hydroponic system is crucial for optimizing both plant growth and energy efficiency. The observed trends highlight the importance of continuous monitoring and adjustment to maintain ideal growing conditions. Future efforts should focus on balancing the need for precise environmental control with sustainable energy usage, leveraging advancements in IoT and machine learning to achieve this goal. The methodology for connecting and operating relay switches with a Node MCU for smart irrigation applications was offered as its conclusion. The outcomes showed that the suggested technique was successful in meeting the goals, including the setup of the hardware, software configuration, relay control logic, integration of sensors and actuators, testing, and safety considerations. In conclusion, "AgroInnovation: Cutting-Edge Technologies Shaping the Future of Farming through IoT" highlights the transformative potential of IoT in agriculture. By integrating advanced sensors, predictive analytics, and smart algorithms, IoT enhances efficiency, sustainability, and productivity. Our study showcases significant power consumption reductions via innovative algorithms, underscoring the importance of energy-efficient practices. As IoT continues to evolve, its application in farming will drive precision agriculture, optimize resource use, and ensure food security. Embracing these technologies is imperative for addressing global agricultural challenges and fostering a sustainable future.

Declaration Statements.

Availability of Data and Materials The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Funding The authors would like to confirm that there was no funding received for the above work.

Conflict of Interest The authors would like to confirm that there is no conflict of interest.

References

1. Shetty, H.M., Pai, K., Mallya, N.: Fully automated hydroponics system for smart farming. Int. J. Eng. Manuf. **11**(4), 33–41 (2021)
2. Khan, S., Purohit, A., Vadsaria, N.: Hydroponics: current and future state of the art in farming. J. Plant Nutr. **44**(10), 1515–1538 (2021). https://doi.org/10.1080/01904167.2020.1860217
3. Alipio, M.I., Dela Cruz, A.E.M., Doria, J.D.A., Fruto, R.M.S.: A smart hydroponics farming system using exact inference in Bayesian network. In: 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), Nagoya, Japan, pp. 1–5 (2017). https://doi.org/10.1109/GCCE.2017.8229470
4. Ramakrishnam Raju, S.V.S., Dappuri, B., Ravi Kiran Varma, P., Yachamaneni, M., Marlene Grace Verghese, D., Mishra, M.K.: Design and implementation of smart hydroponics farming using IoT-Based AI controller with mobile application system. J. Nanomater. **2022**, 12. Article ID 4435591 (2022). https://doi.org/10.1155/2022/4435591
5. Sisyanto, R.E.N., Kurniawan, N.B.: Hydroponic smart farming using cyber physical social system with telegram messenger. In: 2017 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, pp. 239–245 (2017). https://doi.org/10.1109/ICITSI.2017.8267950
6. Sharma, N., Acharya, S., Kumar, K., Singh, N., Chaurasia, O.P.: Hydroponics as an advanced technique for vegetable production: an overview. J. Soil Water Conserv. **17**(4), 364–371 (2018)
7. Saraswathi, D., Manibharathy, P., Gokulnath, R., Sureshkumar, E. and Karthikeyan, K.: Automation of hydroponics green house farming using IoT. In: 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), pp. 1–4. IEEE (2018)
8. Swain, A., Chatterjee, S., Vishwanath, M.: Hydroponics in vegetable crops: A review. Pharma Innov. J. **10**(6), 629–634 (2021)
9. Bugbee, B.: Nutrient management in recirculating hydroponic culture. In: South Pacific Soilless Culture Conference-SPSCC, pp. 99–112 (2003)
10. Chen, Z., et al.: Hydroponic root mats for wastewater treatment-a review. Environ. Sci. Pollut. Res. **23**, 15911–15928 (2016)
11. Alipio, M.I., et al.: On the design of nutrient film technique hydroponics farm for smart agriculture. Eng. Agric., Environ. Food **12**(3), 315–324 (2019)
12. Nurhasan, U., et al.: Implementation IoT in System monitoring hydroponic plant water circulation and control. Int. J. Eng. Technol. **7**(4), 122–126 (2018)
13. Adidrana, D., Surantha, N.: Hydroponic nutrient control system based on internet of things and K-nearest neighbors. In: International Conference on Computer, Control, Informatics and its Applications, pp. 166–171. IEEE (2019)
14. Mehra, M., et al.: IoT based hydroponics system using deep neural networks. Comput. Electron. Agric. **155**, 473–486 (2018)

15. Dudwadkar, A., et al.: Automated hydroponics with remote monitoring and control using IoT. Int. J. Eng. Res. Technol. **9**(6) 2020
16. Gertphol, S., Chulaka, P., Changmai, T.: Predictive models for lettuce quality from internet of things-based hydroponic farm. In: 22nd International Computer Science and Engineering Conference, pp. 1–5. IEEE (2018)
17. Pitakphongmetha, J., et al.: Internet of things for planting in smart farm hydroponics style. In: International Computer Science and Engineering Conference, pp. 1–5. IEEE (2016)
18. Changmai, T., Gertphol, S., Chulak, P.: Smart hydroponic lettuce farm using internet of things. In: 10th International Conference on Knowledge and Smart Technology, pp. 231–236. IEEE (2018)
19. Palande, V., Zaheer, A., George, K.: Fully automated hydroponic system for indoor plant growth. Proc. Comput. Sci. **129**, 482–488 (2018)
20. Lata, S., et al.: Development of greenhouse-application-specific wireless sensor node and graphical user interface. Int. J. Inf. Technol. **15**(1), 59–68 (2023)
21. Mishra, R., Devi, M.P., Gupta, S., Kumar, D.: Smart farming beyond boundaries: global Collaboration and Innovation. In: Agriculture 4.0, pp. 391–405. CRC Press (2024)
22. Mehedi, I.M., Hanif, M.S., Bilal, M., Vellingiri, M.T., Palaniswamy, T.: Remote sensing and decision support system applications in precision agriculture: challenges and possibilities. IEEE Access (2024)
23. Matthew, U.O., Kazaure, J.S., Ebong, G.N., Ndukwu, C.C., Nwanakwaugwu, A.C., Nwamouh, U.C.: Smart IoT rural agricultural innovations for food security through digital technology transfer: agro-industrial sector reform. Glob. Res. Environ. Sustain. **1**(7), 30–50 (2023)
24. Rahman, M.M., Sarkar, S.R., Ghosh, K., Khatun, M., Chowdhury, N., Akter, T., Islam, M.A.: Role of government and private sector on agribusiness development: comparison between the USA and Bangladesh. ORGANIZE: J. Econ., Manag. Financ. **3**(2), 71–90 (2024)
25. Ivanova, D.K.: Centre of excellence and the challenges for the pedagogical specialists from professional education. Recommended for publication by the Publishing Center of Varna University of Management (Minutes No 1 of June 15, 2024), p. 189 (2024)

Unpacking the Knowing-Doing Gap Among Pre-service Teachers Regarding Artificial Intelligence in Education

Sanskriti Rawat[1], Robert Linton Tavis Ashton Bell[1], Prakasha G S[1(✉)], and Sergey Grigoriev[2]

[1] Christ University, Bangalore, India
prakasha.gs@christuniversity.in
[2] Moscow City Pedagogical University, Moscow, Russia

Abstract. Artificial Intelligence is rapidly transforming educational landscapes, yet the preparedness of future educators to engage with AI remains underexplored. This study examines the knowledge-practice gap among pre-service teachers in India through a data-driven analysis of their knowledge, attitudes, and practices regarding the integration of AI in education. A stratified random sample of 223 PSTs from diverse teacher education institutions participated in the study. Data were collected using an adapted KAP questionnaire and analysed using correlation techniques to examine the interrelationships among knowledge, attitudes, and practices. Findings reveal that while PSTs reported moderate to high levels of knowledge and generally positive attitudes toward AI, these did not consistently translate into effective classroom practices. Significant correlations emerged among KAP components, highlighting strong associations between knowledge and attitude, but weaker links with practice. These results underscore the persistence of a knowing-doing gap, suggesting that awareness and favourable perceptions alone are insufficient to foster meaningful AI integration. The study contributes to the growing field of educational data mining by providing empirical insights into the role of AI in teacher preparation. It emphasises the urgent need for targeted training and structured capacity-building programs within teacher education institutions to bridge the gap between what pre-service teachers know about artificial intelligence and how they implement it in practice.

Keywords: Artificial intelligence · Correlation analysis · Educational data mining · India · Knowing-doing gap · Pre-service teachers

1 Introduction

1.1 Shift to the Humachine Era

The world has undergone rapid changes in response to societal needs and demands. Technological advancements have impacted every domain, including education. Interestingly, as humans and machines converge for the greater societal good, they form a symbiotic

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 561–575, 2027.
https://doi.org/10.1007/978-3-032-27834-0_51

pair, known as the humachine concept, to amplify capabilities and address complex problems without erasing human identity [1]. Cutting-edge technological advancements, such as Artificial Intelligence (AI), Mixed Reality (MR), and the Internet of Things (IoT), are integral to meeting the needs and demands of both current and future generations [2]. Digital technologies, such as AI, facilitate innovative pedagogies, including flipped classrooms, gamification, and immersive learning environments (e.g., virtual, augmented, and mixed reality) [3–6]. These approaches foster student engagement and creativity, which are crucial for developing skills in a rapidly evolving world [7]. Sustainable practices also allow educators to navigate their role in the classroom amidst societal changes and advancements. Integrating technology in education aligns with the United Nations Sustainable Development Goals, which state, "[e]ducation for all has always been an integral part of the sustainable development agenda" (United Nations n.d.). This aligns with Goal 4: Quality Education (2021). By equipping teachers and learners with the necessary 21st-century skills and knowledge, technology helps address global challenges and fosters social and economic development [8].

1.2 Artificial Intelligence in Education

AI has been around for decades, but its usage has been relatively recent, particularly since the COVID-19 pandemic. It seeped into teaching-learning and administrative processes and has boomed over time. AI-powered tools are now being integrated into education, particularly in the teaching and learning process, by educators and learners, indicating a significant leap into the AI era [9]. Moreover, AI is now utilised as a supporting tool in the teaching and learning process [10]. Technologies like generative AI (GenAI) redefine education, focusing on personalised, interactive, and on-demand learning. GenAI explores the potential of educational sustainability as virtual tutors, aiding in real-time examination and assessment and being utilised as teaching-learning resources by automating administrative tasks [11]. GenAI tools prompt more effective pedagogical strategies, benefiting future generations of writers, researchers and educators [12]. Furthermore, the practical applications of LLMs are diverse, including improving translations, generating search terms, offering writing suggestions, and providing creative prompts for students [13, 14]. AI-powered tools can also provide personalised experiences and step-by-step guidance based on the learner's understanding and preferred pace [15]. Moreover, the round-the-clock availability and accessibility of AI-powered tools, unlike traditional classrooms, allow learners to ask questions, explore topics, or get assistance at their convenience [16]. GenAI tools, when integrated into education, have the potential to resonate with each learner, and in cases where ethical and sustainable approaches are considered, can play the role of a peer or a mentor [17]. Such a paradigm shift requires educators to be aware of the latest educational technology and its effective implementation in teaching [18].

1.3 Pre-service Teachers and AI

Pre-service teachers are at the forefront of integrating the newest technological innovations as educators in training, and in the context of sustainable education practices, this is of utmost importance. In developing countries like India, it becomes increasingly

important to make pre-service teachers aware of the roles AI can play in education and contextualise them with the necessary ethical understandings, so that the education system as a whole does not lose its primary purpose: empowering learners to be leaders of tomorrow. Pre-service teachers are situated in a vital position to adopt technology, particularly AI-powered tools, within their teaching practices [19], and according to Fidan and Gencel [20], "pre-service teachers need to enhance their learning performance and intrinsic motivation to equip themselves with the requirements of the 21st century" (page 7). To reach their full potential, pre-service teachers need to be familiar with AI tools, maintain a positive attitude towards them, and incorporate them into their teaching and learning scenarios while considering the ethical implications associated with the use of AI in general [21]. Exploring various AI tools can give pre-service teachers the edge in creating engaging classroom materials and enhancing their own professional development as future educators by improving learner engagement and facilitating collaborative learning experiences [22]. The Knowledge, Attitude, and Practices (KAP) framework is a cornerstone for assessing people's learning experiences, perceptions, and behaviours in diverse fields, including education. It was adapted to evaluate the interplay between pre-service teachers' knowledge of these tools, their attitudes toward adoption, and their practical applications in classrooms [21]. With the rising use of GenAI tools like ChatGPT by teachers and students worldwide, the KAP framework is relevant in understanding how educators and students adapt to technological advancements in teaching and learning. The present study addresses this gap by investigating the knowing-doing gap among Indian pre-service teachers concerning the use of AI in education. Using a stratified random sample of 223 PSTs from diverse Teacher Education Institutions (TEIs), the research employs correlation analysis to explore the relationships among knowledge, attitudes, and practices. By deriving AI-driven, data-based insights, this study contributes to the ongoing discourse on the ethical and effective integration of AI in teacher education, providing evidence-based directions for capacity building and curriculum design.

Research Objectives

- To measure the knowledge, attitude, and practices of pre-service teachers regarding AI in education.
- To explore the interrelationship among knowledge, attitude, and practices of pre-service teachers regarding the use of AI.
- To explore if there is a knowing-doing gap among pre-service teachers regarding AI in education.
- To recommend strategies for teacher educators to bridge the knowing-doing gap among pre-service teachers regarding AI in education.

Research Questions

- What are pre-service teachers' measures of knowledge, attitude and practice regarding AI in education?
- What are the relationships among knowledge, attitude, and practices of pre-service teachers towards AI?
- Does a knowing-doing gap exist among pre-service teachers regarding AI in education?

- What strategies can teacher educators adopt to bridge the gap between knowing and doing among pre-service teachers regarding AI in education?

2 Theoretical Framework

The present study is grounded in a constructivist stance, a learning theory rooted in the works of Piaget and Vygotsky [23]. Constructivism emphasises learners' active construction of knowledge through interaction with their environment and collaboration. The current study explores how pre-service teachers perceive the potential of AI to facilitate active learning, collaboration, and personalised learning experiences for their learners. This aligns with the constructivist view of learning as an active and collaborative process. Furthermore, the study builds upon this foundation using Davis' Technology Acceptance Model (TAM) [24], which explores factors that influence an individual's decision to adopt new technologies. TAM posits that the belief that AI can enhance teaching and learning, as well as the belief that AI tools are easy to learn and use, are crucial determinants of technology adoption. Thus, the study's focus on pre-service teachers' engagement with AI tools is framed appropriately. More specifically, the study is grounded in the Knowledge, Attitude, and Practice (KAP) framework, which maps to typical learning taxonomies associated with the cognitive [25], affective [25], and psychomotor [26–29] domains. The KAP model is used to understand how knowledge acquisition influences attitude and ultimately leads to behavioural practices. The model assumes a linear progression, i.e., knowledge shapes attitudes, which in turn influences practices. Robledo et al. [21] used the KAP framework to investigate the knowledge, attitude and practices of pre-service teachers in the Philippines towards the use of ChatGPT (a generative AI tool). For the current study, the KAP framework is relevant as it provides a structured lens to assess not only how much pre-service teachers know about AI tools but also how positively they perceive these tools and how they implement them in their teaching practice.

3 Materials and Methods

This study employed a quantitative, cross-sectional, and descriptive research design to assess the KAP of pre-service teachers toward generative AI tools in education. The target population for this study consisted of pre-service teachers enrolled in teacher education programs across various states in India. The study employed a stratified random sampling technique to recruit participants who met specific inclusion criteria: pre-service teachers actively enrolled in teacher education programs and with prior exposure to AI tools. To ensure that participants met these criteria, the researchers conducted targeted outreach through professional and social platforms, including LinkedIn, WhatsApp, and Facebook. Only those who confirmed their eligibility were reached out to participate in the study, ensuring relevance and alignment with the research objectives. A total of 223 pre-service teachers participated in the study, enrolled in teacher education institutions across various states in India, including New Delhi, Haryana, Karnataka, Uttar Pradesh, Tamil Nadu, and a few other states, such as Gujarat, Assam, and Himachal Pradesh. These participants were selected from various teacher education institutions, including

central, state, and private universities. The sample size of 223 was deemed sufficient for both exploratory and correlational analysis. The present study adapted the original KAP-CQ39 instrument developed by Robledo et al. [21], which was designed to assess pre-service teachers' knowledge, attitude, and practices regarding the use of ChatGPT in the Philippines. In their final validated version, the authors removed one item, resulting in a 39-item scale. However, for the current study, all 40 original items were retained to explore how the full instrument functions within the Indian context of teacher education. Unlike the scoring structure used by Robledo et al. [21], the present study employed a Likert scale to assess the KAP of pre-service teachers and obtain their self-perceptions regarding the use of GenAI tools in education. Changes were made to the item statements in accordance with the study's requirements. Additionally, all negatively worded items were reverse-coded during analysis to maintain consistency in scoring direction. The adaptation process was carried out with the permission of the original authors. The final adapted questionnaire, comprising 40 items across knowledge, attitude, and practice dimensions, is included in the Appendix. The adapted survey comprised 40 items categorised under three dimensions of the KAP framework related to the use of generative AI tools in teaching and learning. Each item was measured on a 5-point Likert scale, ranging from strongly disagree (1) to strongly agree (5), with items that were negatively-worded (items 2, 3, 6, 7, 22, 24–26, 28 and 33) reverse-coded. In addition to the KAP items, the survey also gathered basic demographic data, including age, gender, and type of institution (state, central, or private), to explore potential influences on the KAP scores. The study was approved by the Ethics Committee of the researchers' university, ensuring adherence to ethical research standards. Before participation, informed consent was obtained from all respondents. Participants were assured of the anonymity and confidentiality of their responses, and that participation was voluntary. All responses were anonymised, and data were stored securely on a password-protected device accessible only to the researchers. Participants completed the form independently, and no personally identifiable information was collected. The final dataset comprised responses from 223 participants (n = 223).

4 Results and Findings

4.1 Demographic Details of the Sample

The study sample consisted of 223 pre-service teachers from teacher education institutions across various states in India, including New Delhi, Haryana, Karnataka, Uttar Pradesh, Tamil Nadu, Gujarat, Assam, and Himachal Pradesh. In terms of gender, the majority of participants were female (n = 189, 84.8%), followed by male participants (n = 33, 14.8%), with one participant (0.4%) opting not to disclose their gender. Furthermore, the age distribution revealed that the largest group of participants (n = 132, 59.2%) fell within the 21–25 years age range. Smaller proportions were aged 26–30 years (n = 20, 9.0%), 31–35 years (n = 17, 7.6%), 36–40 years (n = 9, 4.0%), 41–45 years (n = 7, 3.1%), 46–50 years (n = 12, 5.4%), and above 50 years (n = 26, 11.7%). Lastly, regarding the type of institution, participants were primarily from state universities (n = 95, 42.6%), followed by private universities (n = 70, 31.4%), central universities (n = 30, 13.5%), deemed-to-be universities (n = 17, 7.6%), and autonomous institutions

(n = 11, 4.9%). Table 1 presents the distribution of the sample by age, gender, and type of institution.

Table 1. Gender, age and type of institution-wise distribution of the sample.

Sl. No.	Parameter	Groups	Distribution (%)
1	Gender	Male	14.8
		Female	84.8
		Preferred not to say	0.4
2	Age	21–25 years	59.2
		26–30 years	9
		31–35 years	7.6
		36–40 years	4
		41–45 years	3.1
		46–50 years	5.4
		More than 50 years	11.7
3	Type of institution/university	State	42.6
		Private	31.4
		Central	13.5
		Deemed to be	7.6
		Autonomous	4.9

4.2 Validity of the KAP Questionnaire

To examine the internal consistency and construct validity of the adapted KAP instrument, the researchers conducted reliability analysis followed by exploratory factor analysis (EFA) in SPSS version 29.0v. Firstly, the researchers reverse-coded the items that were worded negatively to maintain consistency within the entire framework. The reliability analysis was then carried out using the 40 items of the instrument. Table 2 shows the result of the Cronbach's alpha test for the 40-item KAP questionnaire.

Table 2. Cronbach's alpha test for KAP questionnaire.

Reliability statistics		
Cronbach's alpha	Cronbach's alpha based on standardised items	N of items
0.865	0.878	40

It is evident from Table 2 that the analysis yielded a Cronbach's alpha of 0.865, indicating high internal consistency among the items. The standardised alpha was slightly

higher at 0.878, confirming that the instrument maintained a high level of internal reliability even after accounting for differences in item variance and scaling. These findings suggest that the adapted KAP questionnaire is a reliable tool for measuring pre-service teachers' knowledge, attitudes, and practices related to generative AI in the Indian context. The researchers also conducted Cronbach's Alpha tests for the sub-scales of knowledge (items 1–15), attitude (items 16–30), and practice (items 31–40). Table 3 presents the Cronbach's Alpha test results for the three subscales: knowledge, attitude, and practice.

Table 3. Cronbach's alpha test for knowledge, attitude and practice.

Subscale	Cronbach's alpha	Cronbach's alpha based on standardised items	N of items
Knowledge	0.772	0.792	15
Attitude	0.819	0.824	15
Practice	0.574	0.600	10

The knowledge subscale (15 items) demonstrated good internal consistency with an alpha value of 0.772. The attitude subscale (15 items) yielded a Cronbach's alpha of 0.819, indicating strong internal consistency. The practice subscale (10 items) had a Cronbach's alpha of 0.584, which is considered moderate for exploratory research but indicates the need for further refinement of items within this dimension. As the overall reliability of the 40 items was found to be excellent, the researchers retained all the items. Moreover, these subscale-specific findings align with the theoretical structure of the KAP framework and provide evidence for the instrument's internal reliability.

4.3 Descriptive and Inferential Statistics

To measure the average knowledge, attitude and practice score of the participants, descriptive statistics were used. Table 4 presents the descriptive statistics for the mean scores of Knowledge (K), Attitude (A), and Practice (P) among the participants. The overall means were consistent across components: Knowledge ($M = 0.71$, $SD = 0.07$), Attitude ($M = 0.71$, $SD = 0.08$), and Practice ($M = 0.70$, $SD = 0.08$). These indicate moderate to high awareness and positive dispositions toward AI integration, with some variability in actual classroom practice.

Table 4. Descriptive statistics for knowledge, attitude, and practice ($N = 223$).

Component	Mean	SD	Min.	Max.
Knowledge (K)	0.71	0.07	0.45	0.87
Attitude (A)	0.71	0.08	0.48	0.89
Practice (P)	0.70	0.08	0.42	0.90

The slightly lower mean in Practice suggests that while PSTs understand and value AI's role in education, they face challenges in applying it effectively-indicating a knowing-doing gap consistent with prior literature [20] (Selwyn 2022). Next, to explore the relationships among the three dimensions of the KAP framework—knowledge, attitude, and practice—the researchers conducted a Pearson correlation analysis, the results of which are presented in Table 5.

Table 5. Pearson correlation analysis for KAP.

	KAv	AAv	PAv
KAv	Pearson correlation	1	
	Sig. (2-tailed)		
AAv	Pearson correlation	0.482**	1
	Sig. (2-tailed)	<0.001	
PAv	Pearson correlation	0.430**	0.339**
	Sig. (2-tailed)	<0.001	<0.001

** Correlation is significant at the 0.01 level (2-tailed).

The results from Table 5 revealed that knowledge was moderately and positively correlated with attitude ($r = 0.482$, $p < 0.001$) and practice ($r = 0.430$, $p < 0.001$). Additionally, attitude was positively correlated with practice ($r = 0.339$, $p < 0.001$). All correlations were statistically significant at the 0.01 level (2-tailed). These findings provide empirical support for the interconnected nature of the KAP framework, suggesting that higher levels of knowledge about generative AI tools are associated with more favourable attitudes and greater practical implementation among pre-service teachers.

4.4 Discussion

The present study aimed to examine the KAP of Indian pre-service teachers regarding the integration of AI in education and to identify the persistence of a knowing-doing gap between conceptual understanding of AI and its classroom application. Rather than focusing on psychometric validation, the focus was on interpreting the interrelationships among the three KAP components and their implications for teacher education. Descriptive results revealed comparable mean scores across the three domains: Knowledge ($M = 0.71$, $SD = 0.07$), Attitude ($M = 0.71$, $SD = 0.08$), and Practice ($M = 0.70$, $SD = 0.08$), suggesting that pre-service teachers possess moderate to high readiness in all areas. However, the marginally lower mean and greater variability in practice scores (range = 0.42–0.90) indicate inconsistencies in translating theoretical knowledge and positive perceptions into classroom implementation. This numerical pattern, though subtle, reveals that "knowing" and "feeling positive" about AI do not necessarily equate to "doing" AI in practice. At the correlational level, knowledge was found to be significantly associated with attitude ($r = 0.482$, $p < 0.001$) and practices ($r = 0.430$, $p < 0.001$), although with different magnitudes of association. The stronger knowledge–attitude relationship reflects that conceptual understanding shapes beliefs and openness toward AI. On the

other hand, the comparatively weaker knowledge-practice association reinforces the existence of a gap between knowing and doing. These findings align with global teacher education research [30], which shows that awareness and enthusiasm rarely translate directly into pedagogical action without scaffolding, training, or practical exposure. The weaker practice outcomes in this study can be attributed to structural and contextual barriers. Many teacher education institutions in India continue to face uneven access to AI-integrated resources, limited practicum opportunities, and insufficient faculty modelling of AI use. Such disparities, especially between urban and rural, public and private colleges, amplify the gap between what teachers know and what they can do. The finding resonates with research studies that argue that technological proficiency alone cannot drive innovation without systemic readiness [31]. From the standpoint of attitude, the findings suggest that pre-service teachers' optimism about AI's educational value coexists with concerns regarding its ethical, technical, and pedagogical implications. They express enthusiasm for AI's potential to enhance learning personalisation but remain cautious about issues of bias, data privacy, and authenticity. This conundrum highlights that attitude acts as a mediating, rather than a direct, driver of practice. Globally, similar KAP distributions and correlations have been reported. For instance, Bae and colleagues [32] found that pre-service teachers across different contexts exhibit high awareness and interest but limited classroom enactment, often constrained by pedagogical confidence and institutional culture. The current findings extend this global pattern to the Indian context, situating the knowing-doing gap as a systemic issue rather than an individual shortfall. The findings highlight the need for pre-service teachers to carefully consider whether AI diminishes authentic human learning and teacher-student relationships. Future educators must carefully balance AI with learner-centred, socio-emotional pedagogies [33, 34]. The next challenge is the lack of ethical consideration while using AI, which is a potential hurdle to achieving quality education. Pre-service teachers need to implement ethical practices while using AI in the classrooms to maintain academic integrity [35]. Lastly, there are challenges of accessibility and inequity. Not all educational institutions have access to AI tools, resulting in disparities and barriers that hinder pre-service teachers from implementing AI technologies, which in turn affects their attitudes and practices [36]. Therefore, pre-service teachers have an important role to play in not only being aware of AI but also being mindful, rational, and welcoming of its potential to foster quality education. To address this, it is essential for teacher education programs to move beyond the dissemination of information to the integration of experiential learning. Embedding AI literacy modules with hands-on microteaching, simulation, and project-based learning can help translate conceptual readiness into applied competence [37, 38]. Moreover, faculty modelling of AI use during instruction can motivate responsible and ethical application [39]. Structured reflection on AI ethics, bias, and human-machine boundaries should complement technical skills to ensure balanced development of cognitive, affective, and behavioural dimensions [40, 41]. Hence, the study empirically confirms a persistent knowing-doing gap among Indian pre-service teachers regarding AI in education. While knowledge ($M = 0.71$) and attitude ($M = 0.71$) remain strong, practice ($M = 0.70$) lags slightly behind, indicating the need for pedagogical scaffolding, institutional investment, and contextually relevant teacher training. Bridging this divide will require teacher educators to consciously integrate AI into practicum and curricular

design, enabling future teachers not only to be aware of AI but also to competently and ethically use it in practice [42]. The findings of this study underscore that pre-service teachers possess foundational knowledge and positive attitudes toward AI, yet these do not consistently translate into practical classroom use. This knowing-doing gap indicates that cognitive readiness alone is insufficient for pedagogical transformation. Therefore, teacher educators must act as mediators, converting awareness into situated competence through structured pedagogical design, experiential exposure, and reflective practice. Researchers henceforth suggest the following strategies for teacher educators to bridge the knowing-doing gap persisting among pre-service teachers regarding AI use.

- **Integrate AI pedagogy into foundational courses**

AI should not be treated as an isolated technical topic, but rather be embedded within pedagogy, assessment, and curriculum courses in teacher education. Modules could focus on AI-driven lesson planning, assessment feedback tools, and adaptive learning platforms. By doing so, pre-service teachers begin to perceive AI not as a separate discipline but as a pedagogical partner that can augment instruction. Studies, such as those by George [43], have demonstrated that the explicit integration of AI tools within core pedagogy courses can lead to more confident and ethical classroom experimentation. Moreover, using AI-enabled learning management systems within teacher education can not only help to identify knowledge gaps, but also focus on the knowledge-practice gaps of the learners (pre-service teachers) [44].

- **Use practice-based learning and microteaching**

Bridging the knowing-doing gap requires hands-on environments. Teacher educators may implement AI microteaching sessions, where pre-service teachers design and deliver short lessons using AI tools (e.g., chatbots, adaptive quiz generators, or text-to-image platforms) [45]. Moreover, guided peer and faculty feedback then helps them reflect on affordances and ethical limitations. Such hands-on experiences can increase both confidence and critical awareness among pre-service teachers related to AI, thus shifting their attitudes from mere curiosity to conscious effort.

- **Model responsible AI use through faculty practice**

Faculty modelling has a significant influence on student teacher behaviour. Teacher educators should demonstrate AI integration in their own instruction, such as using AI to summarise readings, co-create rubrics, or support differentiated instruction. This normalises ethical and contextually grounded AI use. Zhang et al. [46] emphasise that observational learning, when grounded in transparency about data bias and algorithmic limits, promotes reflective and responsible adoption [46].

- **Embed ethical reflection and data literacy**

Ethical competence is a crucial bridge between knowledge and responsible practice. Structured activities; such as case analysis on algorithmic bias, discussions on data privacy, or evaluating AI-generated lesson content, can enhance ethical and critical AI literacy. These reflections help future teachers internalise why and how AI should

serve pedagogical rather than efficiency goals, ensuring that integration remains learner-centred and humanistic [47].

- **Strengthen institutional and policy support**

The knowing-doing gap can also be bridged through systemic support. Teacher education institutions must ensure access to AI tools, provide training for faculty, and design curricula that allow AI experimentation without fear of error. Collaboration between universities, EdTech platforms, and policy bodies can create AI teaching-learning communities: safe, experimental spaces for pre-service teachers to explore applications before entering formal classrooms.

- **Encourage continuous professional development**

Ultimately, the dynamic nature of AI necessitates a lifelong learning approach. Teacher educators can promote professional learning communities, AI clubs, and peer mentoring programs that extend engagement beyond coursework. Such initiatives foster an ongoing culture of experimentation and reflection, narrowing the knowing-doing gap over time.

5 Conclusion

The present study examined the knowledge, attitudes, and practices of Indian pre-service teachers regarding AI in education, uncovering a persistent knowing-doing gap between conceptual understanding and practical application. Although participants demonstrated moderate to high levels of knowledge and positive attitudes, they did not consistently report effective classroom practices of AI. The correlation analysis further highlighted that while knowledge and attitude are significantly linked, their relationship with practice remains comparatively weaker, reinforcing the gap between what teachers know about AI and how they enact it in pedagogical contexts. These findings extend global evidence on teacher readiness for AI integration by situating it within the Indian teacher education system, where disparities in access, digital infrastructure, and institutional support often constrain practical experimentation. The study underscores that awareness alone is insufficient to drive transformation; meaningful change requires systemic and pedagogical interventions. Bridging the knowledge-practice gap requires teacher educators to model AI-integrated pedagogy, embed experiential and reflective learning within pre-service curricula, and provide continuous opportunities for ethical and hands-on engagement with AI tools. By doing so, teacher education institutions can empower future teachers not only to understand the concept of AI but also to apply it confidently, critically, and responsibly in real classrooms. This study contributes to the ever-evolving discourse on AI in education by emphasising the translation of knowledge into practice: a crucial step toward preparing an educator workforce that is both technologically.

Conflict of Interest. Authors have no competing interests.

Funding Statement. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Consent to Participate All participants provided informed consent prior to their participation in the study. Participation was voluntary, and respondents could withdraw at any time without penalty.

Consent to Publish No identifying information of participants has been included in this manuscript.

Ethics Statement. The study was approved by the Institutional Review Board of CHRIST (Deemed to be University), Bengaluru. All procedures performed were in accordance with the ethical standards of the institution.

References

1. Mourtzis, D., Angelopoulos, J., Panopoulos, N.: The future of the human-machine interface (HMI) in society 5.0. Futur. Internet **15**, 162 (2023). https://doi.org/10.3390/fi15050162
2. Chirume, S.: Sustainable professional development of primary school mathematics teachers in Zimbabwe through philosophy of education 5.0: challenges and prospects. Sumer. J. Soc. Sci. **3**(12), 150–161 (2020). https://www.researchgate.net/profile/Silvanos-Chirume/publication/347449505_Sustainable_Professional_Development_of_Primary_School_Mathematics_Teachers_in_Zimbabwe_Through_Philosophy_of_Education_50_Challenges_and_Prospects/links/624170e957084c718b7065af/Sustainable-Professional-Development-of-Primary-School-Mathematics-Teachers-in-Zimbabwe-Through-Philosophy-of-Education-50-Challenges-and-Prospects.pdf
3. Asad, M.M., Naz, A., Churi, P., Tahanzadeh, M.M.: Virtual reality as pedagogical tool to enhance experiential learning: a systematic literature review. Educ. Res. Int. **2021**(1), 1–17 (2021). https://doi.org/10.1155/2021/7061623
4. Bao, Y., Li, B.: A preliminary study on graduate student instructors' exploration, perception, and use of ChatGPT. Int. J. Comput.-Assist. Lang. Learn. Teach. **13**(1), 1–23 (2023). https://doi.org/10.4018/ijcallt.332873. https://www.igi-global.com/article/a-preliminary-study-on-graduate-student-instructors-exploration-perception-and-use-of-chatgpt/332873
5. Prakasha, G.S., Rawat, S., Basak, I., Mathai, S.: Gamification in Teacher Education: Teacher Educators and Trainee Teachers' Perspectives in India, pp. 132–162 (2024). https://scholar.google.com/citations?view_op=view_citation&hl=en&user=F0hWtN0AAAAJ&authuser=1&citation_for_view=F0hWtN0AAAAJ:9yKSN-GCB0IC
6. Prakasha, G.S., Thamaraikannan, S., Lapina, M., Sanskriti, R.: Pre-service and in-service teachers' perceptions of using virtual reality tools in teaching. In: Lecture Notes in Networks and Systems, pp. 340–355. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-78776-8_34
7. Prabadevi, B. et al.: Metaverse for Industry 5.0 in NextG Communications: Potential Applications and Future Challenges (2023). arXiv [cs.CY]. http://arxiv.org/abs/2308.02677
8. Dzvimbo, M.A., Mashizha, T.M., Zhanda, K., Mawonde, A.: Promoting sustainable development goals: role of higher education institutions in climate and disaster management in Zimbabwe. Jamba **14**(1), 1206 (2022). https://doi.org/10.4102/jamba.v14i1.1206
9. Tzoneva, I.: Benefits and challenges in using AI-powered educational tools. In: Education and New Developments 2023, vol. 2 (2023). https://doi.org/10.36315/2023v2end079
10. Reiss, M.J.: The use of AI in education: practicalities and ethical considerations. Lond. Rev. Educ. **19**(1) (2021). https://doi.org/10.14324/lre.19.1.05
11. Morosanu, G.-A., Rata, L.-A., Geru, M.: Outcomes of large language models and artificial intelligence in education. Didact. Danub. **3**(1), 30–52 (2023). https://dj.univ-danubius.ro/index.php/DD/article/view/2473. Accessed 22 Feb 2024

12. Alqahtani, T. et al.: The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res. Soc. Adm. Pharm. **19**(8), 1236–1242 (2023). https://doi.org/10.1016/j.sapharm.2023.05.016
13. Morrison, R.: Large Language Models and Text Generators: An Overview for Educators (2022). http://files.eric.ed.gov/fulltext/ED622163.pdf. Accessed 22 Feb 2024
14. Hsiao, Y.-P., Klijn, N., Chiu, M.-S.: Developing a framework to re-design writing assignment assessment for the era of large language models. Learn. Disabil. Res. Pract. **9**(2), 148–158 (2023). https://doi.org/10.1080/23735082.2023.2257234
15. Uzumcu, O., Acilmis, H.: Do innovative teachers use AI-powered tools more interactively? A study in the context of diffusion of innovation theory. Technol., Knowl. Learn. **29**(2), 1109–1128 (2024). https://doi.org/10.1007/s10758-023-09687-1. Accessed 20 Jun 2024
16. Alasadi, E.A., Baiz, C.R.: Generative AI in education and research: opportunities, concerns, and solutions. J. Chem. Educ. **100**(8), 2965–2971 (2023). https://doi.org/10.1021/acs.jchemed.3c00323
17. Ahmad, S., Umirzakova, S., Mujtaba, G., Amin, M.S., Whangbo, T.: Education 5.0: Requirements, Enabling Technologies, and Future Directions (2023). arXiv [cs.CY]. http://arxiv.org/abs/2307.15846
18. Mastour, H., Khoshgoftar, Z., Ahmady, S., Sohrabi, S.: Implementing an internet of things-based learning platform into medical education: addressing educational challenges with innovative solutions. Med. Teach. **47**(8), 1262–1265 (2025). https://doi.org/10.1080/0142159X.2024.2445036
19. Sun, N., Saleh, S.: Mapping the landscape of AI implementation in STEM and STEAM education: a bibliometric analysis. In: Communications in Computer and Information Science, pp. 99–111. Springer Nature Singapore, Singapore (2024). https://doi.org/10.1007/978-981-96-0205-6_7
20. Fidan, M., Gencel, N.: Supporting the instructional videos with chatbot and peer feedback mechanisms in online learning: the effects on learning performance and intrinsic motivation. J. Educ. Comput. Res. **60**(7), 1716–1741 (2022). https://doi.org/10.1177/07356331221077901
21. Robledo, D.A.R. et al.: Development and validation of a survey instrument on knowledge, attitude, and practices (KAP) regarding the educational use of ChatGPT among preservice teachers in the Philippines. Int. J. Inf. Educ. Technol. **13**(10), 1582–1590 (2023). https://doi.org/10.18178/ijiet.2023.13.10.1965. http://www.ijiet.org/show-193-2575-1.html
22. Hastomo, T., Mandasari, B., Widiati, U.: Scrutinizing Indonesian pre-service teachers' technological knowledge in utilizing AI-powered tools. J. Stat. Educ. **18**(4), 1572–1581 (2024). https://doi.org/10.11591/edulearn.v18i4.21644. https://edulearn.intelektual.org/index.php/EduLearn/article/view/21644. Accessed 21 Jun 2024
23. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes, vol. 86 (1978). https://books.google.com/books?hl=en&lr=&id=RxjjUefze_oC&oi=fnd&pg=PA1&dq=Vygotsky,+L.+S.,+%26+Cole,+M.+(1978).+Mind+in+Society:+Development+of+Higher+Psychological+Processes.+Harvard+University+Press.&ots=okA1YYtZfp&sig=u9pe6RC_zzXWhkWzPK4Nak_aKoQ
24. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**(3), 319 (1989). https://doi.org/10.2307/249008
25. Krathwohl, D.R.: Taxonomy of educational objectives. Affective Domain (1973)
26. Armstrong, R.J.: Developing and Writing Behavioral Objectives (1970)
27. Goldberger, M., Moyer, S.: A schema for classifying educational objectives in the psychomotor domain. Quest **34**(2), 134–142 (1982). https://doi.org/10.1080/00336297.1982.10483772
28. Harrow, A.J.: A Taxonomy of the Psychomotor Domain: A Guide for Developing Behavioral Objectives (1972). https://cir.nii.ac.jp/crid/1130282270225137664

29. Simpson, E.: Educational Objectives in the Psychomotor Domain (1985)
30. Nazaretsky, T., Ariely, M., Cukurova, M., Alexandron, G.: Teachers' trust in AI-powered educational technology and a professional development program to improve it. Br. J. Educ. Technol. (2022). https://doi.org/10.1111/bjet.13232
31. Webster, A., Gardner, J.: Aligning technology and institutional readiness: the adoption of innovation. Technol. Anal. Strat. Manag. **31**(10), 1229–1241 (2019). https://doi.org/10.1080/09537325.2019.1601694
32. Bae, H., Hur, J., Park, J., Choi, G.W., Moon, J.: Pre-service teachers' dual perspectives on generative AI: benefits, challenges, and integrating into teaching and learning. Online Learn. **28**(3), 131–156 (2024). https://doi.org/10.24059/olj.v28i3.4543
33. Palmquist, A., Sigurdardottir, H.D.I., Myhre, H.: Exploring interfaces and implications for integrating social-emotional competencies into AI literacy for education: a narrative review. J. Comput. Educ. (2025). https://doi.org/10.1007/s40692-025-00354-1
34. Pratyusha, Y.B., Varghese, B.: A learner-cantered educational landscape: era of education 5.0 and disruptive technologies. In: Mitigating Learner Disadvantages in Teaching and Learning, pp. 159–180. IGI Global Scientific Publishing (2025). https://www.igi-global.com/chapter/era-of-education-50/368634
35. Dimitriadou, E., Lanitis, A.: A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. Smart Learn. Environ. **10**(1), 12 (2023). https://doi.org/10.1186/s40561-023-00231-3
36. Mahapatra, H., Ingole, S.B., Singh, D., Dutt, A., Gupta, R., Chauhan, B.: Artificial intelligence for sustainable education in India: problems and prospects. Libr. Prog.-Libr. Sci., Inf. Technol. Comput. **44**(3) (2024). https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=09701052&AN=180918440&h=m5EySD2TIHuiUorm8nBxaHrPsbEgGbPQDJLqFogxsNsqv2iXlAZaljWvR0h23Xqog1cChFJ77I8VZeoknsLRCQ%3D%3D&crl=c
37. Napier, P., Wada, T.: Bridging the skills gap: the design education landscape with AI and experiential learning. Des. Manag. J. **19**(1), 19–30 (2024). https://doi.org/10.1111/dmj.12094
38. Alexandrowicz, V.: Artificial intelligence integration in teacher education: navigating benefits, challenges, and transformative pedagogy. J. Educ. Learn. (2024). https://doi.org/10.5539/jel.v13n6p346
39. Chan, M.M., Rosales, M., Hernandez-Rizzardini, R., Amado-Salvatierra, H.R.: Ethical AI in education: a proposed model for responsible integration. In: 2025 IEEE Global Engineering Education Conference (EDUCON), pp. 1–7. IEEE (2025). https://doi.org/10.1109/educon62633.2025.11016451
40. Aliman, N.: Hybrid Cognitive-Affective Strategies for AI Safety (2020). https://research-portal.uu.nl/en/publications/hybrid-cognitive-affective-strategies-for-ai-safety
41. Magliocca, P., Canestrino, R., Carayannis, E.G., Gagliardi, A.R.: Understanding human–technology interaction: evolving boundaries. Eur. J. Innov. Manag. (2024). https://doi.org/10.1108/ejim-04-2024-0341
42. Ismail, A., Aliu, A., Ibrahim, M., Sulaiman, A.: Preparing teachers of the future in the era of artificial intelligence. J. Artif. Intell., Mach. Learn. Neural Netw. **44**, 31–41 (2024). https://www.researchgate.net/profile/Akilu-Ismail/publication/381345794_Preparing_Teachers_of_the_Future_in_the_Era_of_Artificial_Intelligence/links/669e6b4902e9686cd118235c/Preparing-Teachers-of-the-Future-in-the-Era-of-Artificial-Intelligence.pdf
43. Shaji George, A.: Preparing Students for an AI-Driven World: Rethinking Curriculum and Pedagogy in the Age of Artificial Intelligence. Undefined [AI], vol. 1, no. 2, pp. 112–136 (2023). https://doi.org/10.5281/ZENODO.10245675. Accessed 06 Oct 2025
44. Afolabi, I.Y., Ugah, J.O., Igwe, J.S., Nwali, M.E.: Development of an AI-driven adaptive learning management system using data analytics. Int. J. Sci. Res. Mod. Sci. Technol. **4**(7), 52–63 (2025). https://doi.org/10.59828/ijsrmst.v4i7.349. Accessed 07 Oct 2025

45. Konakbayeva, U., Baltasheva, P., Kuanysheva, B.: Artificial intelligence in microteaching lesson study: enhancing pre-service teachers' confidence and instructional quality. Educ. Process: Int. J. (EDUPIJ) **15**(1), 1–13 (2025). https://www.ceeol.com/search/article-detail?id=1339435. Accessed 07 Oct 2025
46. Zhang, C., Schießl, J., Plößl, L., Hofmann, F., Gläser-Zikuda, M.: Acceptance of artificial intelligence among pre-service teachers: a multigroup analysis. Int. J. Educ. Technol. High. Educ. **20**(1), 49 (2023). https://doi.org/10.1186/s41239-023-00420-7
47. Zou, D., Xie, H., Kohnke, L.: Navigating the future: establishing a framework for educators' pedagogic artificial intelligence competence. Eur. J. Educ. **60**(2) (2025). https://doi.org/10.1111/ejed.70117

Application of Artificial Intelligence in Ensuring the Safety of Aircraft Flights by Building a Typical Architecture of the Neural Network of an Aviation Enterprise

Nina Romancheva[1](✉), Leonid Firsov[2], Dmitriy Melnik[2], and Ivan Rumyantsev[1]

[1] The Moscow State Technical University of Civil Aviation, 123449 Moscow, Russia
n.romancheva@mstuca.ru

[2] Ecolibri Joint Stock Company, Leningradsky Prospekt 37, 125167 Moscow, Russia

Abstract. The possibility of using artificial intelligence in ensuring the safety of aircraft flights at an aviation enterprise is being considered. Disclosed is an algorithm for creating a neural network architecture capable of identifying critical elements of an aviation enterprise based on available data on hazard factors, which in turn enables to take effective corrective measures to prevent aviation events in advance. Initial data on hazard factors for a neural network are formed based on Fuzzy Sets theory, with continuous monitoring of the safety status of flights of an aviation enterprise in conditions of insufficient statistics.

Keywords: Safety · Aviation enterprise · Unclear level of risk · Hazards

1 Introduction

Based on the national strategy for the development of artificial intelligence for the period up to 2030, the use of artificial intelligence technologies in the sectors of the economy is general ("cross-cutting") in nature and contributes to the creation of conditions for improving the efficiency and the formation of fundamentally new areas of activity of economic entities, including by increasing security, including forecasting risks and adverse events.

Currently, in the new economic conditions, the business of the Russian Federation is trying to direct more resources to fulfill its current tasks, as well as to expand projects and develop new sales markets. This is facilitated by the extension of the moratorium on business inspections until the end of 2024. In this regard, issues of ensuring flight safety in civil aviation of the Russian Federation may fade into the background, in violation of the concept of "safety space," according to which a civil aviation organization must find a balance between the desired level of productivity/profitability and the necessary level of protection through a safety risk control system [1].

Thus, the effective provision of organizational and financial resources to ensure aircraft flights is an essential component of the aviation industry in the context of sanctions

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 576–589, 2027.
https://doi.org/10.1007/978-3-032-27834-0_52

challenges and restrictions. In this regard, artificial intelligence technologies, as a new form of promising management methods, can programmatically provide intellectual support for decision-making aimed at effectively ensuring the safety of aircraft flights.

2 Setting the Task of Building a Typical Architecture of the Neural Network of an Aviation Enterprise Aimed at Ensuring Flight Safety

In accordance with international standards of the International Civil Aviation Organization (ICAO) [2], safety management is based on risk assessment, which provides for the application of predictive methods for analyzing safety data. At the same time, in accordance with the international standard "Risk Management. Principles and Guidance" [3–5], risk is the effect of uncertainty on targets, with uncertainty being a state of insufficiency, even partial, information, understanding or knowledge regarding an event, its consequences or its possibility. Therefore, the risk assessment is a kind of forecast in conditions of uncertainty. At the same time, in civil aviation of the Russian Federation, risk is understood as the estimated probability and severity of the consequences of the manifestation of hazard factors [6, 7]. But in fact, in highly reliable (fault-tolerant) systems, events by which the probability can be calculated in advance are very rare (occur with a probability of "almost—zero"). In this regard, it seems necessary to introduce a new definition of "risk" based on a fuzzy event of "possibility" with a fuzzy scale of significance (importance, danger), in accordance with the theory of fuzzy sets "Fuzzy Sets" and the use of linguistic variables. This approach makes it possible to investigate conflict situations taking into account the uncertainty (inertia) of information about rare events in highly reliable systems, with the help of a new mathematical object in the form of a new indicator $\overline{R}$—"integral risk" (with a cover: "^"), which sets the level of danger based on preventive (proactive) forecast and analysis of the functional properties of the system, given by a certain set of initial properties of the system with the symbol Σ_0 [8].

From the point of view of the presence of two main management systems at civil aviation service providers of the Russian Federation, namely the quality management system (QMS) and the flight safety management system (FRMS), uncertainty can be considered within the boundaries of these systems if measurable quality indicators and flight safety indicators are introduced. At the same time, within the framework of the integration of the QMS and DBMS, three areas of uncertainty of the service provider can be distinguished, or in our case, the aviation enterprise [9–11].

$$\dot{C}_A = M_Q \cup M_S \boxtimes \left(M_Q, M_S\right) \Rightarrow \left\{U_Q, U_s, U_\Sigma | \Sigma_0\right\}, \tag{1}$$

where—$\dot{C}_A$—the system of the aviation enterprise, M_Q—a set of quality indicators, M_S—a set of flight safety indicators, U_Q—the area of uncertainty within the boundaries of the QMS, U_S—the area of uncertainty within the boundaries of the SAP, U_Σ—the area of uncertainty of the integrated quality and safety management system, Σ_0—a set of properties of the production system of the aviation enterprise.

In the considered context, quality indicators establish the degree of implementation of procedures, flight safety indicators are determined in the form of deviations (errors, violations) from the established procedures.

The first area of uncertainty—the calculated certainty lies within the boundaries of the SUK, since production processes and procedures are established at the aviation enterprise before the start of production activities and, for example, for an aircraft operator who plans to carry out commercial air transportation of passengers and cargo, the Federal Air Transport Agency assesses the mandatory requirements established in FAP-10 [12, 13] before issuing an operator certificate. I.e., compliance with FAP-10 requirements and is a data source for the first type of uncertainty, the adjustment of the characteristics of which is within the power of the organization that plans to start commercial air transportation of passengers and cargo.

The second type of uncertainty—uncounted uncertainty lies within the boundaries of the DBMS since deviations of this type of uncertainty occur during or after the performance of the production activities of the aviation enterprise. Deviations that could cause an accident or a serious aviation incident are investigated to develop and implement measures to prevent similar events in the future.

The third type of uncertainty—vaguely calculated uncertainty lies within the boundaries of an integrated quality and safety management system. Values in this type of uncertainty can be obtained by normalizing quality indicators and flight safety indicators in the range from 0 to 1, which corresponds to Fuzzy Sets theory. At the same time, linguistic variables for quality indicators are measured in the form of the degree of achievement of procedures and have the following values:

not reached: (0–0.2)
partially achieved: (0.2–0.5)
mostly achieved: (0.5–0.7)
fully achieved: (0.7–1).

Linguistic variables for flight safety indicators are measured as the number of deviations per accounting period and have the following values:

very small: (0–0.3)
small: (0.3–0.5)
many: (0.5–0.8)
very much: (0.8–0.1).

At the same time, the accounting period is the period established by the aviation company to obtain initial data during continuous monitoring and measurement of quality and safety indicators.

Thus, two sets of indicators are formed for the study of vaguely measurable uncertainty, in which it is possible to determine the critical values of quality indicators of the aviation company's procedures and processes to timely develop and implement corrective measures that can eliminate an adverse event in further production activities, such as a disaster, accident, emergency, serious aviation incident.

Critical values can be detected due to the relationship (feedback) between quality indicators and flight safety indicators in their complex structure, which an aircraft safety researcher can form in the form of a digital twin of an integrated quality and flight safety management system using a neural network.

In this regard, it is possible to generate input data for a neural network that can perform data manipulation functions in a vaguely measurable uncertainty.

The scientific task of building the architecture of the neural network of the integrated quality and safety management system of an aviation enterprise will be to compile a unique set of internal layers of the neural network with the ability to process fuzzy data obtained during continuous monitoring and measurement of quality indicators (Q) and flight safety indicators (S) (see Fig. 1).

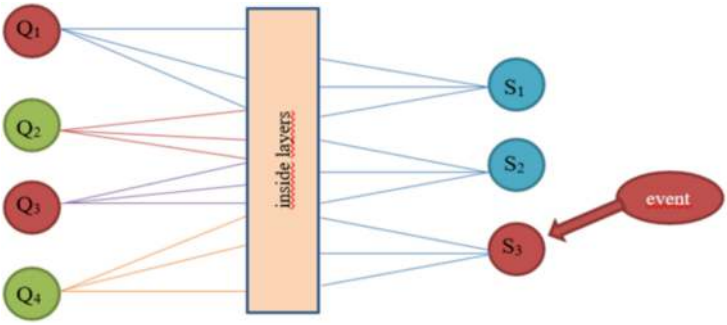


Fig. 1. Diagram of the neural network of the integrated quality and safety management system of the aviation enterprise.

In this case, at a certain stage (with sufficient data), the neural network will form a unique system of weights (fuzzy multi-criteria performance indicators) of the integrated quality and safety management system within the boundaries of fuzzy uncertainty (2).

$$\dot{K}_{\Sigma} = f_{00}\left(M_Q, M_S | \Sigma_0, \Phi_0, \Phi_1\right), \tag{2}$$

where $\dot{K}_{\Sigma}$—a set of multi-criteria performance indicators, Φ_0—quality-related hazards, Φ_1—safety-related hazards, Σ_0—conditions of the production environment of the aviation enterprise.

I.e., for (2) it is required to create an operation algorithm in which the neural network interactively points to critical elements (from a plurality of quality indicators).

$$\widehat{K_{\Sigma}} \Rightarrow \tilde{K} \rightarrow \widehat{K_*} = \widetilde{f_{01}}\left(\tilde{K}\right) \rightarrow M_Q^*, \tag{3}$$

where $\tilde{K}$ is a vaguely measurable performance measure, $\widehat{K_*}$ is a vaguely measurable performance measure indicative of a critical element, M_Q^* is a set of critical elements.

Thus, directly in measurable certainty, within the boundaries of the QMS, critical elements from quality indicators are determined, which in turn allows timely development of corrective measures to prevent an aviation event at an aviation enterprise.

3 Generation of Initial Data for Building a Typical Architecture of the Neural Network of an Aviation Enterprise Introduction

Initial data for monitoring and measurement of quality indicators and flight safety indicators at the aviation enterprise may be generated because of continuous monitoring of indicators, which may include (but not be limited to) such activities as:

internal and external audits (checks) of production processes.
analysis of aviation personnel competencies.
inspection checks of subdivisions, aircraft.
materials of flight information (objective control).
monitoring of changes in indicators.
aviation equipment reliability programs.
investigation materials of aviation events.
materials of mandatory and voluntary communications.
review of safety information.
scenario analysis and modeling.

In accordance with the mandatory requirements in the field of civil aviation for an aviation enterprise performing commercial air transportation [14–16], 166 flight safety indicators can be established, and based on Appendix 1 to PRAPI-89 [17], as well as in accordance with the materials of the investigation of aviation events, 72 flight safety indicators can be established from the Archive of Incident and Industrial Accident Investigation Materials of the Federal Air Transport Agency (AMRIPP Rosaviation) (Table 1). In this case, it is advisable to divide the indicators into 5 classes according to functional characteristics (Table 1) [18].

Table 1. Aviation enterprise indicators.

Classes	Class name	Number of indicators, quality indicators (Q)/safety indicators (S)
1 Class	General organization indicators	13/1
2 Class	Indicators of organization of flight operations	60/22
3 Class	Aircraft airworthiness maintenance indicators	32/16
4 Class	Aircraft ground handling performance	39/20
5 Class	External interaction indicators	22/13
		Total: 166/72

To build a typical neural network of an aviation enterprise, indicators from class 2 “Organization of flight work” were taken as a basis.

The values of the indicators were obtained during continuous monitoring of the activities of the aviation enterprise, for 36 months. Thus, 60 quality indicators and 22 flight safety indicators were monitored. The result is a matrix of 82 columns and 36 rows (see Fig. 2).

The values of the indicators were measured in the range from 0 to 1, during the continuous monitoring of quality and safety indicators, and each indicator was assigned values corresponding to the following linguistic variables (Table 2).

Месяцы	Q2.1.1	Q2.1.2	Q2.1.3	Q2.1.4	Q2.1.5	Q2.1.6	Q2.1.7	Q2.1.8	Q2.1.9	Q2.1.10	Q2.2.1	Q2.2.2	S2.2.4	S2.3.1	S2.3.2	S2.3.3
1	0,5	0,2	0,2	0,5	0,5	0,7	0,5	0,5	0,2	0	0	0,5	0,5	0	0	0
2	0,5	0,2	0,2	0,5	0,5	0,7	0,5	0,5	0,2	0	0	0,5	0,5	0,3	0,3	0
3	0,5	1	0,2	0,2	0,2	0,7	0,2	0,5	0,2	0	0	0,5	0,8	0,3	0,3	0
4	0,5	1	0,2	0,5	0,5	0,7	0,2	0,5	0,2	0	0	0,5	0,3	1	0,5	0
5	0,2	0,5	0,2	0,5	0,2	0,7	0,2	0,5	0,2	0,5	0	0,5	0,5	0,5	0,5	0,3
6	0,2	0,5	0,2	0,2	0,2	0,7	0,5	0,5	0,2	0	0	0,5	1	0,5	0,5	0,3
7	0,2	0,5	1	0,2	0,2	0,7	0	0,5	0,2	0	0	0,5	0,8	0,8	0,5	0,3
8	0,2	0,5	0	0,2	0,2	0,7	0	0,5	0,2	0	0	0,5	0,5	0,5	0,3	0
9	0,2	0,2	0	0,2	0,2	0,7	0	0,5	0,2	0	0	0,5	1	0,5	0,3	0
10	0,2	0,2	0	0,2	0,2	0,7	0,2	0,5	0,2	0	0	0,5	0,5	0,3	0,3	0
24	0,2	1	1	0,4	1	0,4	0	0	0,2	0,2	0,3	0,5	0,1	0,4	0,4	0
25	0,5	0,5	1	0,8	1	0,4	0	0	0,2	0,2	0,3	0,5	0,2	0,2	0,5	0
26	0,2	0,5	1	0,8	1	0,4	0	0	0,2	0,2	0,3	0,5	0,1	0,4	0,4	0
27	0,2	0,5	1	0,8	1	0,4	0	0	0,2	0,2	0,3	0,5	0,1	0,4	0,4	0
28	0,2	1	1	0,5	1	0,4	0	0	0,2	0,2	0,3	0,5	0,1	0,4	0,4	0
29	0,2	1	1	0,5	1	0,4	0	0,8	0,2	0,2	0,3	0,5	0,1	0,4	0,6	0
30	0,2	1	1	0,5	1	0,4	0	0,8	0,2	0,2	0,3	0,5	0,2	0,4	0,2	0
31	0,2	1	1	0,5	1	0,4	0	0,8	0,2	0,2	0,3	0,5	0,1	0,5	0,5	0
32	0,2	1	1	0,5	1	0,4	0	0,8	0,2	0,2	0,3	0,5	0,1	0,5	0,5	0
33	0,2	1	1	0,5	1	0,4	0	0,8	0,2	0,2	0,3	0,5	0,3	0,6	0,4	0
34	0,2	1	1	0,5	1	0,4	0	0,5	0,2	0,2	0,3	0,5	0,1	0,4	0,4	0
35	0,2	1	1	0,5	1	0,2	0	0,4	0,2	0,2	0,3	0,5	0	0,6	0,5	0
36	0,2	1	1	0,5	1	0,2	0	0,2	0,2	0,2	0,3	0,5	0,1	0,2	0,6	0
β	14	15	16	17	18	19	20	21	22	23	24	25				

Fig. 2. Matrix of quality and safety indicators.

Table 2. Linguistic variables for quality and safety indicators.

Evaluation of quality indicators (process, procedures) Q	Evaluation of safety indicators, S	Linguistic terms
Not achieved—0–20% achievement (0–0.2)	Number of deviations/month—5 and more than	0.8
Partially achieved—20–50% achievement (0.2–0.5)	Number of deviations/month—3–4	0.5–0.8
Mainly achieved—50–70% achievement (0.5–0.7)	Number of deviations/months—2	0.3–0.5
Fully achieved—70–100% achievement (0.7–1)	Number of deviations/months—1	0–0.3

Note The column "assessment of flight safety indicators" was formed for a "conditional" aviation enterprise operating about 20 aircraft

4 Building and Training a Neural Network with a Teacher

Based on the obtained conditional data (Fig. 2), the neural network [19–21] of the aviation enterprise was built, which is a digital twin of the production system of the aviation enterprise. The neural network was performed in Jupiter using the Pytorch library. The type of neural network problem is a regression problem that can predict based on a selection of objects with different features.

The neural network was "trained" based on initial data from the matrix of quality indicators and flight safety indicators (Fig. 2). At the same time, the values for quality

indicators were input data for the neural network, and flight safety indicators were output data of the neural network. I.e., 60 inputs and 22 outputs of the neural network were installed.

Thus, the neural network architecture was created based on continuous monitoring of the production system of the aviation enterprise for 36 months, thereby establishing internal links between quality indicators and flight safety indicators of the 2nd class of indicators "Organization of flight work."

The resulting neural network consists of an input layer, 4 hidden layers and an output layer. Hidden layers consist of 3 ReLU () and one ELU () activation functions. This architecture was chosen because it can study complex dependencies in data but is not retrained. This is facilitated by non-linear activation functions and a gradual decrease in dimensionality, which allows you to extract higher-level features. The combination of ReLU and ELU avoids the problem of gradient attenuation and provides stable model learning [22–26]. The architecture of the neural network is shown in Fig. 3.

```
<bound method Module.parameters of Sequential(
  (0): Linear(in_features=60, out_features=120, bias=True)
  (1): ReLU()
  (2): Linear(in_features=120, out_features=240, bias=True)
  (3): ReLU()
  (4): Linear(in_features=240, out_features=176, bias=True)
  (5): ReLU()
  (6): Linear(in_features=176, out_features=88, bias=True)
  (7): ReLU()
  (8): Linear(in_features=88, out_features=44, bias=True)
  (9): ELU(alpha=1.0)
  (10): Linear(in_features=44, out_features=22, bias=True)
)>
```

Fig. 3. Neural network architecture.

For learning, the mean absolute error was used as the loss function:

$$l(x, y) = L = \{l_1, \ldots, l_N\}^T, \quad l_n = |x_n - y_n| \tag{4}$$

During experiments, its use led to the most accurate predictions of the model.

RMSprop (root mean square propagation) or RMS propagation was used as an optimizer. RMSprop is an adaptive optimization method that allows you to adjust the learning rate for each parameter based on gradient history [27–30]. This helps to improve model convergence and avoid the problems of too large- or small-scale updates. For regression problems where it is important to minimize prediction error, RMSprop can provide efficient model learning by considering the features of gradients in different parts of the network.

Updating the weights with RMS propagation is done as follows

$$E\left[g^2\right](t) = E\left[g^2\right](t-1) + (1-)\left(\frac{dc}{dw}\right)^2 \tag{5}$$

$$w_{i,j}(t) = w_{i,j}(t-1) - \frac{\eta}{\sqrt{E\left[g^2\right]}} \frac{dc}{dw_{i,j}} \tag{6}$$

where $E\left[g\right]$ is the moving average of the squares of the gradients, $\frac{dc}{dw}$ is the gradient of the cost function by weight, η is the learning rate, β is the moving average parameter (the default value is 0.9 so that the sum of the default gradient values 0.1 on nine mini-packets and −0.9 on tenths is close to zero, and the value η according to experience, is 0.001).

The adaptive learning schedule is shown (see Fig. 4). It shows what the accumulated impulse looks like for different values of the β parameter.

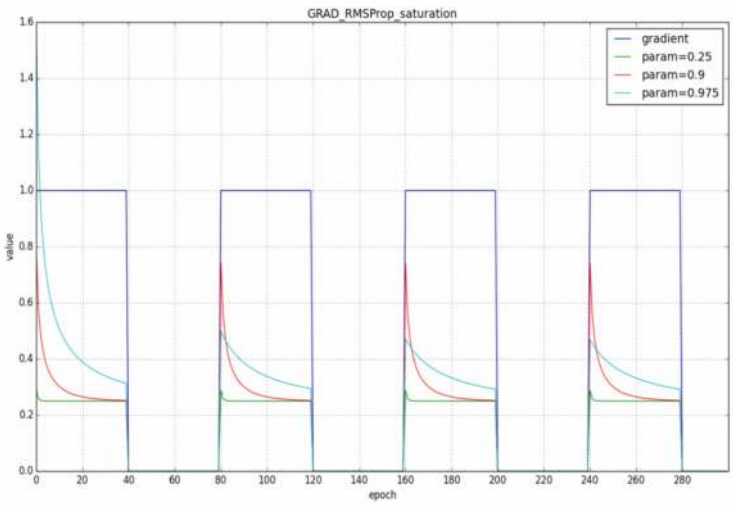


Fig. 4. Adaptive neural network learning schedule.

The x-axis represents epochs, and the y-axis represents gradient values and accumulated values. The accumulated value of the gradient on the epoch is highlighted in blue. Other colors show the value of the Param pulse at different values of the moving average β. So, a frequently updated parameter has a large $E\left[g\right]$, which means a large one. It should be noted that, unlike other adaptive optimization methods, RMSprop does not lead to too small updates due to the gradient square averaged over history [31, 32]. The graph shows that updates do not decrease to zero but reach a certain level. This feature allows the algorithm not to stop at local minima, while maintaining learning speed. During training, gradients were calculated using the backward error propagation method to update the weights.

Thus, the following neural network scheme was formed (see Fig. 5).

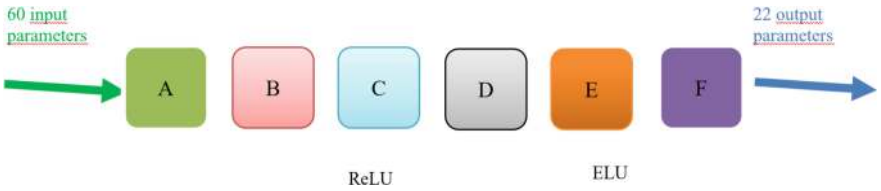


Fig. 5. Neural network diagram.

This neural network can be written as follows:

$$M_Q = \{1, 60\} \Rightarrow \{ReLU\,(A, B, C, D), ELU\,(F, E)\} \Rightarrow M_s E = \{1, 22\} \tag{7}$$

where *A, B, C, D, F, E* are linear layers of the form:

$$y = weight * x + bias, \tag{8}$$

where y is the output of the layer, *weight* is the weight matrix that connects the input and output, x is the input (vector) obtained from the previous layer, or the input for the first layer, *bias* is the shift vector that is added to the output.

The weights of the input and output data of the linear layers with corresponding shear vectors were as follows:

for linear layer A: weight = <120 * 60>, where bias = <120>.
for linear layer B: weight = <240 * 120>, where bias = <240>.
for linear layer C: weight = <176 * 240>, where bias = <176>.
for linear layer D: weight = <88 * 176>, where bias = <88>.
for linear layer E: weight = <44 * 88 >, where bias = <44>.
for linear layer F: weight = <22 * 44>, where bias = <22>.

Thus, based on the results of observations of quality indicators and flight safety for 36 months, the neural network formed a unique set of weights and "shift vectors" of the inner layers, selected by the operator, which in turn turned out to correspond to the structure of the production system of the aviation enterprise.

5 Neural Network Approbation (Unsupervised Neural Network Training)

Approval of the neural network was carried out based on data obtained during an internal audit of the production processes of an aviation enterprise (conditional data). In this case, estimates were obtained for 2 classes of indicators: organization of flight operations. Below are the values of quality indicators normalized in accordance with Table 2, obtained during the internal audit (60 indicators):

0,2 1 0,3 0,2 0,3 0,7 0,2 0,5 0,2 0 0 0,7 0 0 0,9 0,8 0,2 0,3 0 0 0,7 0,9 0 0 0 0 0 0 0 0,5 0,9 0,7 0,9 0 0 1 1 0 0 0 0,6 1 1 0,7 0,8 0,7 0,9 0,1 0,1 0,6 0 0 0 0 0 0 0,9 0 0,9 0,9

LP data were input into the incoming layer of the neural network. On the output layer, 22 values of flight safety indicators were obtained:

A fuzzy risk assessment matrix was used to assess the criticality of the obtained safety outputs (see Fig. 6).

Evaluation of fuzzy risk level ($\hat{R}$) of hazardous events occurrence			Damage category				
			A	B	C	D	E
		multiplier	1	0,8	0,6	0,4	0,2
Degree of danger	1-я	1	1	0,8	0,6	0,4	0,2
	2-я	0,8	0,8	0,64	0,48	0,32	0,16
	3-я	0,6	0,6	0,48	0,36	0,24	0,12
	4-я	0,4	0,4	0,32	0,24	0,16	0,08
	5-я	0,2	0,2	0,16	0,16	0,08	0,04

Fig. 6. Fuzzy risk matrix ally.

At the same time, critical combinations of elements from a variety of quality indicators and flight safety indicators are determined when the values of fuzzy multi-criteria performance indicators lying in an unacceptable fuzzy risk level are exceeded.

Thus, the considered production system of the aviation enterprise has conditions for the occurrence of the following possible events from the flight safety indicators (S):

1. Preliminary preparation for flights is not performed.
2. Deviations (errors) when working with aviation equipment.
3. Not providing designated rest days for flight crew members.

Corrective actions for quality (Q) indicators should be developed for proactive safety management. Therefore, it is necessary to draw up a mathematical relationship in which it was possible to select elements from quality indicators, which are a chain of critical elements recognized before the conditions for a catastrophe arose. This relationship is called the "catastrophe equation" [33–35]. This equation is designed according to the minimum cross-section method—in the form of conjunctions of critical elements β_i of the production system, leading to deviations from established procedures or to the vulnerability of the system under the influence of external factors. In our case, according to the data obtained because of an internal audit of the production processes of an aviation enterprise, the "catastrophe equation" will take the following form (9):

$$\hat{R} \to L_R \Rightarrow U_R = \begin{pmatrix} \beta_2 \wedge \beta_{15} \wedge \beta_{16} \wedge \beta_{22} \wedge \beta_{31} \wedge \beta_{33} \wedge \beta_{36} \wedge \beta_{37} \wedge \beta_{42} \\ \wedge \beta_{43} \wedge \beta_{45} \wedge \beta_{47} \wedge \beta_{48} \wedge \beta_{57} \wedge \beta_{59} \wedge \beta_{60} \end{pmatrix} = 1, \quad (9)$$

where $\hat{R}$—the unclear risk level of the production system, L_R—the structure of the scenario chain, U_R—the disaster condition, β_i—the name of the element of the production system (serial number)—a total of 17 critical elements. Condition $U_R = 1$ means the occurrence of a catastrophe. I.e., out of 60 quality indicators, 17 indicators were identified. After corrective actions, Eq. (9) will take the form (10):

$$\hat{R} \to L_R \Rightarrow U_R = 0 \quad (10)$$

6 Validation of Neural Network Data Processing Results

To check the validity of the results of processing the data of the neural network, all zeros (0) were entered into the input layer (60 inputs) in the first case, i.e., the danger is currently minimal and in the second case all ones (1), i.e., the danger is currently maximum.

In the first case (all zeros), the following data were obtained at the output (22 outputs):

0,3 0 0 0 0 0,2 0,2 0,1 0,3 0,2 0 0 0 0 0 0 0 0 0 0 0,1 0,1

In the second case (all units), the following data were obtained at the output (22 outputs):

0,8 0,1 0,1 0,1 0,6 0,4 0,9 0,5 0,7 0,7 0 0,3 0 0 0,1 0 0 0 0 0 0,5 0,2.

Based on the obtained output values (22 values), it can be concluded that if, according to the results of an internal audit of the production processes of an aviation enterprise, the quality-related hazards are zero, then out of 22 output data, 8 values barely exceed the zero value and according to the matrix for assessing the fuzzy level of risk (Fig. 6) the risk of these 8 hazards is acceptable. Therefore, the aviation company does not need to take any corrective actions in this case. The deviations of 8 values from zero can be explained by the inertia of the production system under consideration since the neural network was initially trained on enough quality-related hazards (17 quality attributes with an unacceptable level of risk).

As for the second set of input data equal to one, out of 22 output data values, 6 safety indicators have an unacceptable level of risk, i.e., exactly 2 times more when training a neural network with a teacher, when 17 quality indicators with an unacceptable level of risk were submitted to the input (when 17 quality attributes with unacceptable risk were submitted, 3 safety attributes with unacceptable risk were obtained).

Thus, the resulting neural network performs its work according to quite logical laws.

7 Conclusion

Thus, a typical neural network architecture has been proposed for an aviation enterprise in which quality indicators are input and flight safety indicators are read at the output. At the same time, the initial data for the neural network were created at the initial stage with the help of a “teacher,” at the next stage the neural network carried out intelligent data processing on its own, using for this purpose input data from quality indicators obtained during an internal audit of production processes. This approach makes it possible to systematize the process of continuous monitoring of indicators of an aviation enterprise, to train a neural network in a continuous mode, while forming a digital twin of the system and, when critical elements appear from quality indicators, to develop corrective measures that can effectively affect the production system in order to exclude an adverse event in further activities of the aviation enterprise, such as a disaster, accident, emergency, serious aviation incident.

References

1. Safety Management Manual (Document 9859), 4th edn. International Civil Aviation Organization (ICAO) (2018)
2. Safety Management Manual. In: Annex 19 to the Convention on International Civil Aviation, 2nd edn. (2016)
3. Risk Management. Principles and Guidelines. In: GOST R ISO 31000-2019, Moscow (2019)
4. Miller, M., Mrusek, B., Herbic, J., Holley, S., Halawi, L.: Implementing an AI fatigue risk management system for aviation maintenance SMS: a technology enhanced critical process human factors safety plan. In: Proceedings of the AHFE, Honolulu, Hawaii (2024). https://www.hawaii.ahfe.org/program.html
5. Araujo, T., et al.: In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. **35**(3), 611–623 (2020). https://doi.org/10.1007/s00146-019-00931-w
6. Howell, K.: How to Monitor and Control Risk in Aviation SMS. http://www.springer.com/lncs. Accessed 23 Sept 2025
7. Poell, T., Nieborg, D.B., Duffy, B.E.: Spaces of negotiation: analyzing platform power in the news industry. Digit. J. 1–19 (2022). https://doi.org/10.1080/21670811.2022.2103011
8. On approval of the rules for the development and application of aircraft flight safety management systems, as well as the collection and analysis of data on hazard and risk factors that pose a threat to the safety of civil aircraft flights, storage, and exchange of these data in accordance with international standards of the International Civil Aviation Organization and recognition of certain acts of the Government of the Russian Federation as invalid. In: Decree of the Government of the Russian Federation of 12.04.2022 No. 642, Moscow (2022)
9. Kuklev, E.A.: Model of the sleeping catastrophe of the reason as an element of the SAP in civil aviation. Problems of security management of complex systems. In: XXXI International Conference 2023, vol. 1, pp. 32–70, Moscow (2023)
10. Balyasnikov, V.V., Kuklev, E.A., Melnik, D.M.: Determination of critical combinations of elements in the field of uncertainty of the states of the aviation system using the example of Sirius-Aero Airlines. Special Issue—Appendix to the Journal Transport Strategy XXI Century **1**(48), 44–50 (2021)
11. Xu, R. et al.: Improved particle swarm optimization based workflow scheduling in cloud-fog environment. In: Daniel, F., Sheng, Q., Motahari, H. (eds.) Business Process Management Workshops. BPM. Lecture Notes in Business Information Processing, vol. 342. Springer (2019). https://doi.org/10.1007/978-3-030-11641-5_27
12. Rukhlinsky, V.M., Kuklev, E.A., Melnik, D.M.: Application of the theory of fuzzy sets in ensuring the safety of flights of civil aviation service providers in conditions of uncertainty of the states of the aviation system. In: Information Document of the Interstate Aviation Committee for the 41st Assembly of the International Civil Aviation (ICAO) A41-WP/72. https://www.icao.int/Meetings/a41/Documents/WP/wp_072_ru.pdf. Accessed 02 Sept 2025
13. Klimenko, A., Barinov, A.: Resource-saving multiobjective task distribution in the fog- and edge-robotics. In: Ronzhin, A., Sadigov, A., Meshcheryakov, R. (eds.) Interactive Collaborative Robotics. ICR 2023. Lecture Notes in Computer Science, vol. 14214. Springer (2023). https://doi.org/10.1007/978-3-031-43111-1_25
14. Subramanya, D., Nyirenda, C.N.: Multi-swarm PSO algorithm for static workflow scheduling in cloud-fog environments. IEEE Access **10**, 117199–117214 (2022). https://doi.org/10.1109/ACCESS.2022.3220239
15. Collaborative Robotics: ICR 2023. Lecture Notes in Computer Science, vol. 14214. Springer (2023). https://doi.org/10.1007/978-3-031-43111-1_25
16. Li, D., Yao, A., Feng, K. et al.: Next Frontiers of Aviation Safety: System-of-Systems Safety Engineering, vol. 52, pp. 262–277 (2025). https://doi.org/10.1016/j.eng.2025.01.002

17. Degas, A., et al.: A survey on artificial intelligence (AI) and eXplainable AI in air traffic management: current trends and development with future research trajectory. Appl. Sci. **12**(3), 1295 (2022). https://doi.org/10.3390/app12031295
18. Ayhan, S., Pesce, J., Comitz, P., Sweet, D., Bliesner, S., Gerberick, G.: Predictive analytics with aviation big data. In: 2013 Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, VA, USA, pp. 1–13 (2013). https://doi.org/10.1109/ICNSurv.2013.6548556
19. Kistan, T., Gardi, A., Sabatini, R.: Machine learning and cognitive ergonomics in air traffic management: recent developments and considerations for certification. Aerospace **5**(4), 103 (2018). ProQuest. https://www.proquest.com/scholarly-journals/machinelearning-cognitive-ergonomics-air-traffic/docview/2582792906/se-2. Accessed 23 Aug 2025. https://doi.org/10.3390/aerospace5040103
20. Mehrabi, N., et al.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021). https://doi.org/10.1145/3457607
21. Gürbüz, F., Özbakir, L., Yapici, H.: Explainable artificial intelligence for aviation safety: a review. Expert Syst. Appl. **213**, 118877 (2023)
22. Amorkor, O., Esmaeil, Z., Roggow, B., Dehghan, N.: Evolution of human factors research in aviation safety: a systematic review and bibliometric analysis of the intellectual structure. J. Saf. Sci. Resil. **7**(1), 100249 (2026). https://doi.org/10.1016/j.jnlssr.2025.100249
23. Zhu, Y.: Formation of flight control of a group of unmanned aerial vehicles based on the algorithm of the multi-agent swarming model. Comput., Telecommun. Control **15**(4), 22–36 (2022)
24. Prüfer, J., Schottmüller, C.: Competing with big data. J. Ind. Econ. **69**(4), 967–1008 (2021)
25. Kagenov, A.M., Kostyushin, K.V., Ismailov, K.K., Kostyushina, N.O., Orlov, S.A., Prokhanov, S.A.: The development of a cloud system for investigation of UAVs aerodynamic characteristics. J. Phys.: Conf. **1488**, 1–5 (2020). https://doi.org/10.1088/1742-6596/1488/1/012017
26. Shatrakov, Y.G., Zatuchny, D.A., Romancheva, N.I., Akinshin, R.N., Avtin, I.V.: Noise Resistance Enhancement in Aircraft Navigation and Connected Systems. Springer Nature Singapore Pte Ltd. (2021). https://doi.org/10.1007/978-981-16-0630-4
27. Sowmya, T., Mary Anita, E.A.: A comprehensive review of AI based intrusion detection system. Meas.: Sens. **28** (2023). https://doi.org/10.1016/j.measen.2023.100827
28. Aldweesh, A., Derhab, A., Emam, A.Z.: Deep learning approaches for anomaly-based intrusion detection systems: a survey, taxonomy, and open issues. Knowl.-Based Syst. **189** (2020). https://doi.org/10.1016/j.knosys.2019.105124
29. Lapina, M., Anita, M., Bagautdinova, A., Lapin, V., Rudenko, M.: Machine learning research methods for identifying inaccurate content. In: Sustainable Computing and Intelligent Systems. Springer Nature Singapore, Singapore (2025). https://doi.org/10.1007/978-981-96-3311-1_16
30. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: theory, method and application. IEEE Trans. Pattern Anal. Mach. Intell. **46**(8), 5362–5383 (2024). https://doi.org/10.1109/TPAMI.2024.3367329
31. Kobanda, A., Maillard, O., Portelas, R.: A continual offline reinforcement learning benc mark for navigation tasks. In: IEEE Conference on Games (CoG), pp. 1–8 (2025)
32. Yang, J., Lin, C.-E.: Autonomous clustering by fast find of mass and distance peaks. IEEE Trans. Pattern Anal. Mach. Intell. **47**(7), 5336–5349 (2025). https://doi.org/10.1109/TPAMI.2025.3535743
33. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 12113–12132 (2023). https://doi.org/10.1109/TPAMI.2023.3275156

34. Hospedales, T., Antoniou, A., Micaelli, P., Storkey, A.: Meta-learning in neural networks: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5149–5169 (2022). https://doi.org/10.1109/TPAMI.2021.3079209
35. Shadab, M., Javed, M.S., Sajid, M., Prasad, M., Lapina, M.A., Babenko, M.: Generating highly nonlinear S-boxes using a hybrid approach with particle swarm optimization In: Nature-Inspired Optimization Algorithms for Cyber-Physical Systems (2025). https://doi.org/10.4018/979-8-3693-6834-3.ch001

Regulatory Documents Machine-Understandability—A Stage in the Formation of the Methodological Basis for Digital Production

Nina Romancheva[1(✉)], Evgeniy Yurkevich[2], and Lidiya Kryukova[2]

[1] The Moscow State Technical University of Civil Aviation, 123449 Moscow, Russia
n.romancheva@mstuca.ru

[2] Institute of Control Sciences of Russian Academy of Science, Profsouznya str. 65, 117997 Moscow, Russia

Abstract. As a basic problem in digital technologies building, the paper proposes to consider the formation of advanced standardization, which determines the most important trends in the design of advanced production lines. In relation to solving the problems of optimizing the regulatory developing technological processes support, the paper proposes the concept of conditions for the application of hybrid intelligence. An approach based on ensuring the machine-understandability of the regulatory documents content as on the effective stage of the digital production formation.

Keywords: Methodological framework · Digital production · Hybrid intelligence · Machine-understandability of content · Regulatory documents · Automation tools · Advanced standardization

1 Introduction

The current intensification of technology development is largely determined by trends focused on meeting the individual requirements of each consumer. In this regard, it is proposed to consider the variety of orders using artificial intelligence (AI) technologies. However, due to the breadth of presentation of even orders of the same type, there is a need for expert solutions. Thus, the problem of the formation of means of hybrid intelligence (GI), combining the capabilities of AI technologies and the experience of decision makers (SCL), is posed.

The development of automation applications in digital industries has shown that existing sets of protocols are often complex, ineffective and in some important cases unrealizable [1, 2]. This makes them difficult to apply and causes resistance from both developers and users focused on other network technologies. At the same time, the spontaneous development of algorithms for creating automation systems in digital industries can lead to the fact that network technologies include functions whose implementation violates the compatibility of new developments [3, 4].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 590–603, 2027.
https://doi.org/10.1007/978-3-032-27834-0_53

The modern problem of ensuring the intensity of technological development presupposes the need to harmonize trends in the creation of digital automation systems for production complexes in various sectors of the country's economic complex. An effective direction in solving such problems when building digital industries can be the creation of a system of advanced standards [5].

We will assume that the methodological basis for the development of digital production is the standardization of software for technological lines. As one of the essential directions for the development of digital automation, considering the breadth of opportunities for improving the technological and organizational mechanisms of the operation of production complexes, we will consider the factor of regulating their information support and its impact on the functional reliability of technological processes.

By functional reliability we will understand the characteristic of the implementation of the technological process, determined by the probability of achieving the goal of its construction. Thus, the task is to identify the conditions under which the system of means of automation of digital production provides maximum functional reliability of technological modes [6].

2 Methodology for Applying Organizational Technologies in the Implementation of New Design Solutions in the Construction of Digital Technological Lines (DTLs)

The specificity of manufacturing processes that utilize digital technologies lies in the diversity of functional capabilities applicable to performing a wide range of operations. A significant characteristic of modern digital manufacturing is the insufficiency of data regarding their practical implementation. Consequently, the pace of technological development may lag the rate of order updates to produce new products [7]. With the aim of introducing the capabilities for adapting digital manufacturing to the specificities of their use at a particular enterprise, this article examines the task of identifying limitations on the organizational technologies employed under conditions of introducing new design solution.

The modern increase in the requirements for the dynamics of the characteristics of production processes has determined the problem of ensuring the functional reliability of technological lines at such production facilities. In this case, it is proposed to consider the inverse problem of optimal control. As a result of its solution, a mechanism must be formed to ensure the adequacy of information interaction of the elements of the technological line with fixed influences of external factors that determine the technological support of the production process. Setting such a problem belongs to the class of incorrect ones, since it assumes a multivariate of possible solutions with similar effects on the technological process at a fixed point in time.

As one of the tools that provide a solution to such a problem, it is proposed to use neural networks (NS). Experience shows that their use can determine the formation of adequate methods and means to ensure regular operation of the technological line [8, 9]. This approach of using artificial intelligence (AI) tools is aimed at convolution of data, which allows algorithmizing the technology of making expert decisions.

Thus, a transition is formed from a purely empirical NS model to a semi-empirical version using hybrid intelligence blocks. As a result, the dimension of the NS model is reduced, its adequacy is increased when using training data sets that are insufficient for traditional NS models. Identification of the characteristics of unstructured variables described by nonlinear functions allows the formation of self-learning mechanisms for NS systems and adaptation to changes in external influences.

The unification of functional tasks solved by expert and AI tools determined the transition from traditional cyber-physical to socio-cyber-physical systems (SCFS). A feature of the SCFS is the introduction of intelligent functions that allow you to quickly correct data on the operation of the human-machine system. Management information support is proposed to be carried out in the format of trainable neural algorithms of socio-cyber-physical self-organization.

This review assumes the use of FMS to control the operation of digital process lines (DTL) [10, 11]. Ensuring the efficiency of CTL control is complicated by the fact that with an increase in the requirements for the accuracy of the implementation of design solutions for manufactured products, the requirements for the functional reliability of the CTL itself increase. In this regard, it is proposed to consider the CTL as a large system in the understanding that it is not the opposite of a "small system," but this term is a characteristic of its description.

Traditionally, systems are defined by values of parameters characterizing relationships between their elements [12]. In this case, it is proposed to characterize a large system with functions defined by transactions between its elements [13].

Due to the multivariate strategies for organizing the technological process of production of each spacecraft, the task of ensuring the reliability of the BA is fundamentally incorrect. To optimize the organizational system of production with restrictions on the available SCFS resources of a specific production complex, the following model is proposed:

$$T_3 = f\left(T_Q, T_S, T_R, (\alpha), (\beta), p_i\right), \tag{1}$$

where T_3 is a comprehensive assessment of the reliability of the technological line to produce functional units, considering the level of completion of the BA assembly processes.

T_Q—reliability factor of SCFS operation, as an integral assessment of the summary of calculations of the dynamics of failures in the technological process (Q).

T_S—efficiency factor of assembly processes organization as integral assessment of importance of parameters of load dynamics of controlled process (S).

T_R—is the stability factor of operations under impacts affecting errors and deviations from the standards as an integral assessment of the importance of the dynamics of process failures (R).

α—characteristic of the importance of the error in the construction of the technological process when using the means of stabilizing the algorithms for assembling the functional units of the BA.

β—characteristic of the importance of error in the construction of the technological process in case of refusal to use means of stabilization of algorithms for assembling functional units of the BA.

p_i—probability of reliable operation (PRO) of the system for compensating the effects of destabilizing factors under the itch strategy in the assembly of OE.

The specific application of software in the operation of the SCPS for technological process management is determined using cybernetic methods. Therefore, such a system is capable of perceiving deviations in the order of operations execution from normative requirements. Monitoring the results of the human factor's influence on process participants on the outcomes of operations serves as an information base for the self-learning of the SCPS and the correction of the entire technological line's operation.

The application of model (1) reveals the conditions for stabilizing the results of BA assembly. As a result, the formulation of the task of optimizing assembly process technology is determined by the constraints (T_Q, T_S, T_R). In practical calculations, the following are used:

- The set Q is characterized by the reliability coefficient of the SCPS operation (T_Q). It determines the PRO of the participants in the work, recorded in a summary of expert assessments of their performance effectiveness.
- The set S characterizes the efficiency of organizing the technological processes of assembly (T_S). It includes parameters of the dynamics of critical loading of the CTL within the time intervals for performing operations, as defined by the project customer.
- The set R is characterized by the coefficient of stability of operation results under influences that cause deviations from normative values (T_R). Based on the monitoring of abnormal situations, it determines the compatibility characteristics of their execution results under the influences of the human factor.

Currently, due to the growing specifics of the software and hardware of technological complexes, the requirements for its information support are increasing. In this review, the implementation of such support is proposed to be considered as a complex information system (IS).

In solving the tasks of ensuring the reliability of the technological complex, based on practical considerations, one of the most significant factors is the integration of IS into the algorithm of the SCFS control and coordination of all production processes. The peculiarity of IE application in the analysis of documentation determining the construction of process lines of the considered class is very high requirements for accuracy and timeliness of operations [14].

These requirements primarily address the resolution of artificial intelligence (AI) challenges, focusing on the need to integrate previously independent systems into a unified framework. In the past, various production tasks—such as planning, management, supply, and sales-were handled autonomously, functioning as isolated "islands" of automation. The new approach demands integrating these tasks alongside diverse technologies, including data processing, text analysis, image recognition, and machine graphics. Essentially, modern complex process-level information systems (ISs) are inherently integrated and interconnected solutions. Furthermore, there is an increasing necessity for machine readability of documentation. Looking ahead, as multiple ISs become

interconnected and collaborate across related processes—such as supply chain organization or fostering relationships with customers and business partners—the feasibility of transitioning to fully machine-readable content will become more evident. For the present, the challenge lies in creating foundational conditions that enable this transition through the optimization and algorithmizing of technological process characteristics, represented in simplified models. A practical model must accommodate the continuous evolution inherent in enterprises operating within a market economy. This includes adapting to regulatory changes that mandate corresponding updates to ISs, which now play a holistic role within organizations and evolve alongside business re-engineering processes. Additionally, within a competitive market economy, stricter constraints are placed on the timeframes for executing technological development and testing, alongside limitations on material and financial resources allocated for such activities. To navigate these conflicting demands, an information model for content representation must possess the following attributes: Interoperability, Portability, Elasticity of systems.

The new approach demands integrating these tasks alongside diverse technologies, including data processing, text analysis, image recognition, and machine graphics. Essentially, modern complex process-level information systems (ISs) are inherently integrated and interconnected solutions. Furthermore, there is an increasing necessity for machine readability of documentation. The trend towards interoperability between systems in related processes, such as the supply chain or strengthening relationships with customers and business partners is becoming more prominent, and the transition to full machine-readability will become more apparent.

The need to consider the specifics of the variety of technological techniques and means of automation of their provision in the process of BA assembly determined the trend of AI development for the use of neural networks. This approach makes it possible to form mechanisms for their construction of more scalable technological processes, where the “curse of dimension” restrained adequacy from representation.

The developed approach to solving systemic problems proposes using promising ultra-precise Convolutional Neural Networks (CNNs) for assembling artificial intelligence tools. The experience of analyzing multidimensional data cubes described in the literature allows for the automatic selection of process structures. This work suggests classifying these structures according to the specifics of their application at a particular production site, which enables the identification and restoration of missing data. It is also possible to use graph neural networks (GNN), which can be used to model the complexity of relationships between data points in irregular grids (for example, process parameters when producing identical functional blocks at different production sites) and build more accurate 3D models of the final product [15]. Generative adversarial networks (GANs) can be used to generate data to supplement training sets and create alternative training scenarios. Recurrent neural networks (RNN, long short-term memory—LSTM, gated recurrent units—GRU) allow you to analyze time series of data for individual functions that make up the process of the technological line. Also, recurrent neural networks can be used to identify promising production sites based on freely available reports [16].

Studying the possibility of using neural networks to digitalize the presentation of assembly production of BA KA and experience in the development of control systems based on neural networks showed that recurrent neural networks (RNN, LSTM, GRU)

can be considered the most expedient for analyzing information management systems with AI in this area. The advantage of using RNN is the ability to store information throughout the entire period of assembly processes, which allows better processing of long sequences of incoming information than other neural networks.

With operational control of the results of each operation, the solution to analyze the data streams taken from sensors in real time, identify failures with the forecast of possible complications minutes before their occurrence, the use of recurrent neural networks (RNN, LSTM, GRU) and transformers (special architecture of deep neural networks—deep neural network—DNN) is promising.

During production, convolutional neural networks (CNNs) can provide data to diagnose the state of each functional unit, which increases the reliability of representing the entire AD design.

An important feature of the organization of such diagnostics and, accordingly, decision-making is the tendency of decision-makers to organize work either at maximum intensity modes, or hypertrophic ally slowed down, with a focus on avoiding risks. A trade-off in the representation of such control modes of a given process is currently achieved using hybrid intelligence tools, where AI tools provide information and operators make decisions.

To solve the problems of machine comprehensibility of content, considering existing solutions, the most expedient can be considered the use of information management systems with AI based on convolutional (CNN) and recurrent (RNN, LSTM, GRU) neural networks.

Physically informed neural networks (PINN) allow integrating fundamental signal filtering equations into the architecture of a neural network model and building more accurate and physically meaningful models of representing the technological process with less training data. To analyze the results of each of the stages of the BA assembly process, it is rational to use convolutional CNN and graph GNN neural networks in digital twins (CD). Experience in assembly operations shows that the use of recurrent neural networks RNN, LSTM, GRU can be considered the most expedient to ensure the safety of the assembly process [16].

Important areas of application for neural networks are the processing of data from sensors installed on auxiliary equipment, detecting anomalies and deviations from its normal mode of operation. Systems built based on neural network technologies predict the likelihood of breakdowns and malfunctions, which allows preventive maintenance and replacement of equipment before it fails.

The presentation of optimal BA assembly schemes by AI-based systems allows considering system data that give a picture, for example, of reducing power consumption or increasing resource efficiency [17]. In solving problems of this type, it can be concluded that the most expedient is to use systems with AI based on convolutional CNN and recurrent neural networks with LSTM architecture.

Modern AI systems based on neural network technologies, in relation to the tasks of building ergatic systems, are divided into the following functional groups:

- means for analysis of technological modes, which are the main source for presenting estimates of production efficiency. The most common types of neural network models for solving the problems of this group are graph GNN and physical informed neural

networks PINN [18], which analyze data streams from SCADA systems, predict the impact of parameter changes on the quality of production results.
- tools that determine the prediction and prevention of complications that may arise during the technological process. The main types of neural network models for solving this class of problems are recurrent RNN, LSTM, GRU neural networks and DNN transformers, which analyze the dynamics of parameter changes in real time to identify early signs of complications during the technological process.

The main neural network models used to work with data in organizing the functioning of technological lines are recurrent neural networks RNN, LSTM, GRU and DNN transformers, which are necessary to analyze data coming from sensors at the input/output of equipment to identify anomalies and predict their residual resource. Multimodal neural networks based on large language models (LLM) are used to combine these SCADA systems. The result of these solutions is a reduction in equipment repair costs, prevention of failures, and optimization of spare parts stocks.

- means of analyzing information that determines the operation of the automated assembly process control system and provides the operator with assistance in making decisions. The main AI model that solves this group of tasks is natural language processing for analyzing text reports, shift logs, instructions for the user to extract knowledge and form recommendations in emergency situations, creating chat bots for quick access to procedures. Computer vision (CV) can be used to analyze video streams while monitoring compliance with safety precautions and detecting leaks (by visual or thermal images).

Further development of specialized regulatory frameworks, defining the algorithmic application of hybrid intelligence methods based on the extensive use of AI-generated databases, will become the methodological foundation for optimizing technological processes, reducing their preparation costs, ensuring high operational safety of equipment, including early accident warning, integrity monitoring, hazardous zone control, reducing the risks of emergency situations, and improving the efficiency of decision-making in technological process management.

It should be noted that the introduction of hybrid intelligence systems based on AI technologies requires ensuring correctness in the interpretation of neural network models with the processing of a large amount of incoming information in a form that is accessible for human perception. A separate factor determining the effectiveness of the use of hybrid intelligence is the organization of specialized training of qualified personnel serving these systems.

The decision on the final choice of specific neural network models for use in systems with AI, allowing to increase the efficiency, ensure the safety and stability of the functioning of information and control systems, is carried out either at the stage of their design, or during the modernization of control systems. Such a solution should consider the uncertainties of external influences on the stability of the functions specified in the specifications. Attempts to eliminate such uncertainties lead to an increase in the set of controlled parameters and the “curse of dimension” in the characterization of the software and hardware used becomes a limiting factor in the organization of given technological processes during a strictly defined time interval.

The specifics of the practical application of the developed approach to building large technological systems confirm the rationality of using big data methodology. As the set of parameters included in the DPL model expands, based on this methodology, the article proposes to identify reserves in CTL projects, as well as vectors for the development of information support to increase the efficiency of technological chains. This approach is since the use of the term big data is usually associated with Google Trends publications since 2011 [18]. Experience in solving the problems of optimizing technological processes shows the promise of its use in ergatic systems characterized by technical and technological parameters under the influence of organizational and/or economic factors, when there is more data than 100 GB (500 GB, 1 TB, etc.). Another example is working with data that cannot be processed in Excel or cannot be processed on the same computer at a given time.

The approach developed in this article is based on the premise that the optimization of production processes is guided by a set of requirements stipulated in regulatory documents. In this context, the assessment of functional reliability of technological processes, through the calculation of the probability of meeting these requirements, relies on the results of analyzing the technical condition of equipment, identifying the least reliable functional blocks based on the characteristics of their operating modes under the influence of external factors.

For example, the use of big data in experimental development of design solutions makes it possible to identify unnecessary tests in assessing the technical condition of processors [18].

Making forecasts based on past data. For example, big data allows you to predict failures in the CTL based on an assessment of the dynamics of such failures during similar production processes or predict the failure of a functional module before this event occurs.

The use of big data in reliable design using digital twins as virtual models makes it possible to predict technological modes. Estimates of the survivability of CTL using the example of digital twins reveal vectors for improving their structural design and optimizing technological operations [19].

The inclusion of big data in the construction of intelligent CTL software reduces the influence of the human factor in automated process control systems and increases the severity of their technical condition control [20].

3 Standardization of Functions as the Basis of the Regulatory Framework of CTL Technologies

The development of algorithms for analyzing process automation systems is inevitably governed by the application of relevant standards. In this context, a set of baseline standards is intended to comprehensively regulate the interaction protocols of production operation system elements, interfaces, and data exchange formats that characterize technological processes [21]. In the presented review, such a set of normative documents is defined as a system profile, and after its approval, the profile becomes a functional standard. The general provisions of functional standardization involve identifying the functions of information systems (IS) and their components, which are fixed as objects

of standardization in the system profile. This allows for the application of standardized design solutions when building IS (analogous to the methods of large-block building construction) to reduce costs and shorten the time frames for creating and implementing IS in conditions of their increasing complexity. In the standard ISO/IEC 100000-1-99 "Information technology. Framework and technology for functional standards. Part 1. Framework and documentation framework," profiles are defined as a subset and/or combination of baseline information technology standards required to implement the specified sets of functions.

The requirements for intensification of spacecraft production determined the organization of simultaneous production of OE units at different sites (at different plants). This situation gives rise to the problem of "inconsistency" of technical and software tools, which leads to unproductive costs associated, among other things, with the problem of retraining personnel when mastering new tools. This is manifested when analyzing the expansion of systems with the need to include new components, as well as when new generations of hardware and software appear, their replication and reuse. The solution to the problem of applying and implementing a methodology that minimizes the "inconsistency" of technical and software tools when creating integrated systems, technologies, and services, reducing the cost of developing these systems, was proposed by foreign specialists working in the field of information technology (IT). This is an open systems methodology. Its main advantages are the preservation of previously invested investments when building IS on various hardware and software platforms, ensuring the interconnection of systems, portability of application software and data [22].

The solution to the problem of applying and implementing a methodology that minimizes the "inconsistency" of technical and software tools when creating integrated systems, technologies, and services, reducing the cost of developing these systems, was proposed by foreign specialists working in the field of information technology (IT). This is an open systems methodology. Its main advantages are the preservation of previously invested investments when building IS on various hardware and software platforms, ensuring the interconnection of systems, portability of application software and data [23].

A new view of open systems is determined by the fact that these features are considered in combination, as interconnected, and implemented in complex. According to IEEE (The Institute of Electrical and Electronics Engineers): "An open system is a comprehensive and consistent set of international IT standards and functional profile standards that specify interfaces, services and supporting formats to ensure the mobility and interoperability of applications, data and personnel" [22].

In the context of solving control problems in digital process automation systems, the CTL norms are proposed to be considered as a paradigm for their development. According to this approach, for each of the process layers of the APCS, an OPC server may be applicable, supplying data to an OPC client at a higher layer or transmitting this data to an adjacent device [23].

The OPC standard is focused on reducing the costs required to build and maintain industrial automation applications. A family of software technologies that provide a single interface for managing automation objects and processes OPC (Open Platform Communications). Many OPC protocols are based on Windows technologies: OLE,

ActiveX, COM/DCOM. OPC protocols such as OPC XML DA and OPC UA are platform independent.

The development of digital production has determined the need to create universal industrial software as a tool for exchanging data with devices from different manufacturers or in accordance with various data exchange protocols. The promise of using OPC is to provide developers of industrial programs with a universal fixed interface (set of functions) for exchanging data with any device. This approach determines the formulation of advanced standardization tasks, in which device developers determine the main trends in creating programs (a set of functions) that implement this interface. The latest version of the OPC DA specification is the OPC UA (Unified Architecture) series of standards, which unify a set of functions for data exchange, event logging, data storage, and data security.

A significant advantage of OPC UA technology is that to support real-time operation, instead of a client/server model, built on this standard, it can use a publisher/subscriber model together with TSN technology. The client/server model works well in the case of a point-to-point connection, but if there are many devices, then there are delays in updating the data. The publisher/subscriber model provides one-to-many and many-to-many connectivity. The server sends its data to the network, and each client can receive it.

An important (often determining) factor affecting the stability of the technological system in uncertain situations is the formation of the transition boundary from one operation to another. In this regard, it is useful to apply the concept of building a CKFS based on the interaction of experts and AI tools. According to the proposed, it is possible to build a technological process of assembling a complex structure because during information interaction each of its participants decides on the implementation of the selected behavior strategy or on its change in connection with the transition of the technological process to a new state.

The transition to digital technologies, that is, a fundamental change in technological solutions, has determined a fundamentally new presentation of the content of regulatory documents. In this regard, it is proposed to use the methodology of functional standardization, when the result of the technological process is regulated with permissible restrictions on the use of resources (material, financial, temporary) [24].

It is proposed to use information in the form of an image as one of the forms of presentation of a functional standard, which allows to unify the regulations of characteristics of a wide variety of production results and restrictions on technological processes. This representation includes the meaning of the regulated process results and their specifics in relation to the technology used. In this case, the target component must be strictly regulated and represented as fixed values of the values of the set R. However, the complete set of parameters defined in model (1) is often not subject to numerical representation. They can be evaluated if their specific assessment is taken for a specific case, and for each participant in the interaction, the assessments will differ [25].

Here, the advantage of regulating the process using a figurative interpretation is manifested: each user perceives the same content in his own way. In the extreme case, their reactions can be diametrically opposite. For example, if in the process of considering a regulatory document, the conditions for regulating characteristics are not clear,

then some users are interested in innovative development of technologies (or design solutions) based on a deeper understanding of the possibilities of implementing regulated functions, although some developers may fall into a state of relaxation. This situation, combined with a wide variety of standards, determined the urgency of raising the problem of building documents using technology based on ensuring their machine comprehensibility.

Due to the wide variety of types of content presented in the standards, to ensure its machine comprehensibility, it is required to form an algorithm for unifying the structure of regulatory documents in the coordinates of the affine space, allowing to interpret the technological process, considering the regulation of impacts on its characteristics, does not change during arbitrary conversion of the set R. As constraints in such a representation of the process plan, means implemented using hybrid intelligence systems are values of sets S and Q.

A significant application of functional standards is the requirement to calculate the probability of error in the use of process stabilization means for the assembly of onboard equipment units. However, obtaining experimental data on the stability of the assembly process for such units is a very costly process, both financially and in terms of time resources [26]. Therefore, stabilization means are not always used in production. At the same time, model (1) introduces into the norms prescribed in functional standards the probabilities of ensuring the stability of the assembly process when using special means (α) and when refraining from their use (β).

A feature of information transmission in ergatic systems is that the throughput of the communication channel between their elements is variable. At the same time, we note that the certainty of the goal setting of users is a condition for the correctness of understanding the advanced content in the regulatory document. For example, to analyze such a document that determines the trend of reorganization of technological processes, two types of systems can be considered:

1. A material and technical system determined by numerical ratios of existing and required characteristics of technological support at each stage of production, as well as restrictions on material and financial resources.
2. An organizational system defined by ways to minimize costs according to expert assessments of specialists of this company.

The analysis of the first type of system involves the consideration of regulatory support only at the level of the traditional presentation of the standard. For example, using a functional cost analysis of the implementation of a technological process, it is possible to obtain numerical values of the estimated costs. However, the correctness of such results obtained only by means of AI is debatable, since the human factor (which determines the “animating” connections of the subjects who implement the technological process) is backbone, and without its consideration, the decision may be incorrect.

In terms of this work, such a ratio is expressed in the similarity of the classification features of the means of implementing the technological process allocated by each of its participants. The probability of coincidence of the most important characteristics highlighted by each of the subjects in question is a significant factor in ensuring the functional reliability of the organizational system.

In this regard, the essence of machine comprehensibility of the proposed system of regulatory documents is the construction of databases systematized based on the results of optimization of their application in the construction of the technological process. We will assume that interlayer functions are considered as services that are not tied to a certain layer of the AI neural network model but may affect more than one layer [27]. Given that the characteristics of communication services are largely determined by the interaction between network design and network control protocols, the task is to introduce methodological provisions into functional standards that define the orientation of orthogonal services towards improving the triad of transmitted data: confidentiality, integrity, and availability.

4 Conclusion

As an implementation of the proposed methodology for advanced standardization, a set of standards "Digital Industry. Unified OPC Architecture," which considers the conceptual vision of the problem of the development of digital transformation of key industries based on standardization in the use of AI technologies in combination with expert decision-making standards. The draft of this series of national standards provides a conceptual view of investing in IT infrastructure and managing digital transformation processes in key industries based on the use of AI technologies. Such technologies are proposed to include the implementation of solutions that provide orientation in the creation of machine-readable means for processing data arrays, considering assessments of the possibilities of their dynamics.

The use of hybrid intelligence tools for the development of regulatory support for digital technological lines provides for the regulation of the functions and semantics of the means of process automation systems. Among other things, the regulation of artificial intelligence tools is proposed to be based on the methodology of object-oriented standardization as based on the formation of norms for describing types/models of the subject area and their interaction, represented by a system of function prototypes that form a hierarchy of technological processes.

In important factor in ensuring machine-readability of content by means of hybrid intelligence is the regulation of expert decision-making mechanisms based on subject-oriented standardization as a method of constructing object-oriented systems in the form of a composition of subjects. This approach allows you to manage the production process itself and implement large software and hardware projects. In fact, we are talking about the progressive organization of information according to the primary semantic criteria: "important/unimportant," "key/details," "parent/child," "single/multiple." Progression makes it possible to move to the next level of detail, which closes the general process of the NS system.

In general, the strategic goals of the Russian Federation, legislatively enshrined by Decree of the President of the Russian Federation of 21.07.2020 No. 474, determined the importance of the leading standards used in the draft set of standards "Digital Industry. Unified OPC Architecture" for methodological and regulatory technical understanding and application of requirements in the field of hardware and software development for data exchange in industrial networks and process control and monitoring systems in the CPU.

References

1. Yurkevich, E.V., Kryukova, L.N.: Problems of regulating the functional reliability of measuring and control instruments in industrial processes. Meas. Equip. (1), 19–23 (2013)
2. Bakumenko, L.P., Minina, E.A.: International index of digital economy and society (I-DESI): trends in the development of digital technologies. Stat. Econ. **2**, 40–54 (2020). https://doi.org/10.21686/2500-3925-2020-2-40-54
3. Belobragin, V.Ya., Zazhazhalkin, A.V., Zvorykina, T.I.: Technical Regulation at the Turn of the Industry 4.0, Moscow (2019)
4. Yurkevich, E.V., Kryukova, L.N.: Problems of regulatory support for the functional reliability of digital production. Reliability **24**(3), 52–60 (2024)
5. Budkin, Yu.V.: The role of standardization in the implementation of import substitution policy. Stand. Qual. **10**, 46–48 (2022)
6. Yurkevich, E., Kryukova, L.: Normative aspects of unification of means of ensuring functional reliability of technological system. In: Proceedings of the 17th International Conference Management of Large-Scale System Development (MLSD), pp. 1–5. IEEE (2024). https://doi.org/10.1109/MLSD61779.2024.10739477
7. Lapina, M., Anita, M., Bagautdinova, A., Lapin, V., Rudenko, M.: Machine learning research methods for identifying inaccurate content. In: Sustainable Computing and Intelligent Systems. Springer Nature Singapore, Singapore (2025). https://doi.org/10.1007/978-981-96-3311-1_16
8. Salimyanova, I.G., Valiakhmetov, I.R.: Innovation Management, 2nd edn. Publishing House of St. Petersburg State University of Economics, St. Petersburg (2023)
9. Lukinova, O.V.: Methodological aspects of information system lifecycle management based on functional standardization tools. Softw. Syst. **4**(29), 27–35 (2016). https://doi.org/10.15827/0236-235X.116.027-035
10. Larin, S.N., Yudinova, V.V., Larina, T.S.: Approach to the creation of an expert system for assessing the quality of innovative program projects of production management. Soc. Hum. Sci. Econ. Issue **3**(45) (2016). https://doi.org/10.18454/IRJ.2016.45.019
11. Rashid, A. B., Kausik, A.K.: AI revolutionizing industries worldwide: a comprehensive overview of its diverse applications Hybrid Adv. **7** (2024). https://doi.org/10.1016/j.hybadv.2024.100277.G
12. Lia, C., Yang, C., Mellerd, C., Lee, J.: Data issues in industrial AI systems: a meta-review and research strategy. Comput. Ind. **175** (2025). https://doi.org/10.1016/j.compind.2025.104361
13. Lia, C., Liua, J., Maa, Y., Ana, X., Lyub, D., Caoa, Y.: A multi-fidelity space-division method is developed to accelerate global optimization. Comput. Ind. **175** (2025). https://doi.org/10.1016/j.advengsoft.2025.103979
14. Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Hu, X.: Data-centric AI: perspectives and challenges. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), pp. 945–948. Society for Industrial and Applied Mathematics (2023)
15. Demir, K.A., Döven, G., Sezen, B.: Industry 5.0 and Human-Robot Co-working. ScienceDirect. Procedia Comput. Sci. Int. http://com/science/article/pii/S1877050919312748. Accessed 15 Sept 2025
16. Castelo-Branco, I., Cruz-Jesus, F., Oliveira, T.: Assessing industry 4.0 readiness in manufacturing: evidence for the European Union. Comput. Ind. **107**, 22–32 (2019). https://doi.org/10.1016/j.compind.2019.01.007
17. Romancheva, N., Krivopalov, D. Yurkevich, E.: Calculation algorithms for the probability of failure-free operation of a large-scale system with a minimum required number of heterogeneous elements of the loaded reserve. In: 14th International Conference Management of Large-Scale System Development (MLSD) 2021, pp. 1–5. IEEE (2021). https://doi.org/10.1109/MLSD52249.2021.9600140

18. Santos, C., Mehrsai, A., Barros, A.C., Araújo, M., Ares, E.: Towards industry 4.0: an overview of European strategic roadmaps. Procedia Manuf. **13**, 972–979 (2017). https://doi.org/10.1016/j.promfg.2017.09.093
19. Romancheva, N.: Multi-agent management architecture under conditions. In: 16th International Conference Management of Large-Scale System Development (MLSD) 2023, pp. 1–4. IEEE (2023). https://doi.org/10.1109/MLSD58227.2023.10304035
20. Magomadov, V.S.: Research potential of industrial internet of things. Int. Res. J. **6**(96) (2020). https://doi.org/10.23670/IRJ.2020.96.6.013
21. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: theory, method and application. In: IEEE Trans. Pattern Anal. Mach. Intell. **46**(8), 5362–5383 (2024). https://doi.org/10.1109/TPAMI.2024.3367329
22. Brooks, R.: An inconvenient truth about AI—AI won't surpass human intelligence anytime soon. IEEE Spectrum (2021). https://mordorintelligence.com/industry-reports/natural-language-processing-market. Accessed 19 Sept 2025
23. Mehrabi, N. et al.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021). https://doi.org/10.1145/3457607
24. Kumar, A., Das, S., Snasel, V.: Efficient three-stage surrogate-assisted differential evolution for expensive optimization problems. Swarm Evol. Comput. **98** (2025). https://doi.org/10.1016/j.swevo.2025.102093
25. Chernova, O.A.: The potential of artificial intelligence in industrial companies' product–service systems. Upravlenets/The Manager. **16**(4), 70–86 (2025). https://doi.org/10.29141/2218-5003-2025-16-4-5
26. Krakovskaya, I.N., Korokoshko, Yu.V., Slushkina, Yu.Yu.: Digital maturity of industrial enterprises: experience estimates. Bulletin of St. Petersburg University. Economy **40**(3), 433–459 (2024). https://doi.org/10.21638/spbu05.2024.305
27. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 12113–12132 (2023). https://doi.org/10.1109/TPAMI.2023.3275156

Information System for Evaluating the Profile of an Information Object

Marina Rudenko(✉), Milada Krapivina, Dmitrii Kuhta, and Andrei Rudenko

V.I. Venadsky Crimean Federal University, Simferopol, Russia
maridigit@gmail.com

Abstract. The widespread dissemination of disinformation and toxic content on the Internet poses a serious problem, which is why the task of evaluating and profiling information objects is particularly relevant. However, metrics obtained using modern artificial intelligence models (e.g., neural networks) are often difficult to interpret, which makes it difficult for experts to use them when analyzing content and making decisions. This paper proposes an information system that combines various content indicators (metrics) obtained by visual, textual, and social (distribution) analysis methods into a unified multimodal assessment of an object's profile. To this end, a set of metrics has been developed to evaluate visual content (e.g., images), text messages, and the nature of information dissemination (on social networks). The results of these analyses are integrated using a production system of rules, allowing a final summary assessment of the information object profile to be formed on the basis of specified threshold values for characteristics. The effectiveness of the proposed approach has been demonstrated in a practical scenario close to real conditions, confirming its practical value and applicability. The resulting assessment is interpretable for specialists and can serve as a basis for subsequent expert verification. Thus, the combination of AI methods and expert rules provides a more transparent and reasonable analysis of content. This increases confidence in the results obtained and expands the possibilities for applying the system in a real information environment. The study was funded by the Russian Science Foundation (RNF) grant (project No. 25–21-20125 "Development of mathematical models and a system for assessing content for extremist content based on text analysis and object detection in images and videos").

Keywords: Disinformation · Toxic content · Multimodal content analysis · Profiling of information objects · Production system of rules · Interpretability · Expert verification · Social networks

1 Introduction

With the growth of digital content and user-generated material on the internet, the task of automatically assessing information objects is becoming increasingly relevant. Every day, huge amounts of text, images, and videos are published, and manual moderation of this content is becoming virtually impossible, as well as highly stressful for experts [1]. Scalable methods based on artificial intelligence (AI) technologies are needed to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 604–618, 2027.
https://doi.org/10.1007/978-3-032-27834-0_54

automatically analyze and filter content [1, 2]. In recent years, such systems have been actively implemented in social media and content sharing platforms to monitor the compliance of published materials with specified rules and criteria.

Modern AI models are already demonstrating impressive results in analyzing various types of content. In the field of text processing, automatic sentiment analysis allows the emotional tone of messages and reviews to be extracted [3]. Deep learning methods are used to effectively identify toxic and offensive comments that pose a threat to communication [4]. In computer vision, deep neural networks have surpassed humans in a number of image classification tasks [5]. Models have been developed for the automatic detection of undesirable visual content—for example, a ResNet-based system for detecting NSFW material is capable of determining the degree of explicitness of images or videos with an accuracy of about 95% [6]. Similarly, neural network classifiers are used for videos to successfully recognize unacceptable scenes. In one study, a combination of a convolutional network and BiLSTM achieved ~95% accuracy in classifying children's video content as safe or inappropriate [7]. These achievements confirm that AI can effectively evaluate the content of both text and multimedia information.

However, content in real-world applications is often multimodal in nature, combining text, images, audio, and video. Therefore, the latest research focuses on combining different modalities for comprehensive evaluation. Approaches are emerging in which text and visual data are analyzed jointly by a single model or a set of models. For example, in brand safety tasks, advertisers predefine undesirable content categories (violence, erotica, extremism, etc.) and use hybrid systems combining machine learning and expert rules to verify compliance with these requirements. Such systems analyze several signals at once—video, audio, text—to classify a video into a particular category according to corporate security standards [1].

The development of large multimodal models (e.g., combining visual and linguistic representations) opens up new possibilities for automatically understanding complex content. However, research shows that approaches relying solely on text analysis using linguistic models miss important visual cues and context [1].

This has stimulated the creation of multisensory algorithms: for example, Zhao et al. developed a deep model that simultaneously takes into account the textual, visual, and temporal features of messages, which significantly improved the accuracy of fake news detection compared to unimodal solutions [8]. Thus, the integration of heterogeneous content data into a unified profiling system is a promising direction in the field of AI

Despite the success in analyzing individual types of data, the problem of integrating the results of different models to produce a single comprehensive assessment of an information object remains unresolved. Standard metrics for the quality of neural networks, such as accuracy, recall, F_1-measure, etc.—are suitable for comparing models with each other [1], but are not very informative in terms of applied content assessment. In other words, a high accuracy rate of a toxic comment classifier indicates the reliability of the model, but does not directly answer the question of whether a particular post or document meets the requirements set by an expert. In practical scenarios, experts formulate the desired content profile as a set of criteria and restrictions, often in an informal way (for example, "the information resource should be informative, without profanity, not

contain scenes of violence, evoke positive emotions in the audience, “ etc.). The individual conclusions of highly specialized models—whether it is the probability of negative text tone or the detection of an object in an image—do not provide a complete picture and are difficult to interpret without context. Moreover, the combination of such diverse metrics is difficult to understand for an expert who does not have in-depth knowledge of AI. In practice, this leads to the need to manually adjust the thresholds for each model and try to reduce them to a single solution. For example, a study on automatic content filtering showed that with a threshold value of 0.5, the Perspective API service detects up to 91% of offensive statements (with $F_1 \approx 0.71$), and when the threshold is lowered to 0.3, it detects up to 94% of unwanted messages ($F_1 \approx 0.87$), at the cost of a slight increase in false positives [9].

Such threshold selection actually requires expert involvement and indicates the need for a more transparent and meaningful human-friendly assessment mechanism that combines various metrics.

This paper proposes an approach to evaluating the profile of an information object, designed to combine the heterogeneous results of AI models and present them in a form that is easy for experts to understand and use. An information system has been developed that aggregates indicators obtained from several models of content analysis of different types (text, visual, audio-visual) into a single set of profile metrics. Each metric reflects a separate aspect of the quality or content of the object and has a clear expert description. For example, metrics of tone, information density, and toxicity level are introduced for texts; for images, the presence of undesirable objects or scenes, aesthetic evaluation, etc. Based on these specific metrics, the system calculates a total integral assessment of the object’s compliance with the specified profile. The key role is played by the profile model specified by the expert—a set of rules and threshold values that determine the desired range for each metric. We have implemented this mechanism through a production rule system: if the object’s metric values fall within the ranges specified by the profile, the system gives a positive final assessment () (otherwise, it gives a negative or intermediate assessment with notes on which criteria are not met). This method makes the process transparent—an expert can track which indicators are used to justify the final decision and how. In addition, profiling based on multiple parameters allows for complex combinations of requirements to be taken into account. Practice shows that enriching the content profile with additional characteristics does indeed improve the quality of decisions: for example, including user engagement statistics in the description of a news article significantly improved the accuracy of the model for predicting reader subscriptions to the publication [10]. Our approach thus combines the strengths of automatic analysis (scalability, objectivity of metrics) with expert knowledge of the subject area embedded in the profile structure.

2 Methods

2.1 Metrics for Evaluating Different Types of Content

2.1.1 Image Evaluation Metrics

We presented them in another work [11]. YOLO detection is a vector sequence of five values. Based on these, a model for analyzing objects in images and videos has been developed. It includes the following parameters (metrics): object area, object focus center, number of objects, maximum confidence, average confidence, maximum area, average area, coverage area, abscissa dispersion, and ordinate dispersion.

Area is a metric for a single detected object. The center of focus, number of objects, maximum confidence, and coverage area can be calculated not only for detected objects of a single class, but also for all objects found (this occurs when multiple images are detected). All others are calculated within a single class.

Object area (S) is the geometric area of an object, which is the product of the width and height of the bounding box of the object in the detected image. The value is relative to the size of the image. The object area clearly shows what part of the entire image is occupied by the detected object.

$$S = \mathrm{w} * \mathrm{h}, \tag{1}$$

where w is the relative width of the object's bounding box,

h is the relative height of the object's bounding box.

Focus center (X, Y)—the average point of the bounding box centers of the objects found; point coordinates—the arithmetic mean of the relative coordinates of the objects' bounding boxes:

$$\mathrm{X} = \frac{\sum_{k=1}^{n} x_k}{n}, \tag{2}$$

x—relative abscissa of the object,

n—number of objects in the class.

$$Y = \frac{\sum_{k=1}^{n} y_k}{n}, \tag{3}$$

y—relative ordinate of the object,

n—number of objects in the class.

If the coordinates of the focus center are equal (or very close) to the coordinates (0.5; 0.5), this means that the objects (object) occupy a central position in the image. The more central the object is in the image, the greater the semantic load it carries for human perception. If we consider the interpretation of the metric value in the subject area of searching for illegal content, we can determine the following: the closer the focus center of the detected object is to the focus center of the entire image, the more likely it is that the object was found by accident and the more "illegal" the image itself is.

Number of objects (n)—the number of objects of one class found or the total number of objects found. This metric can be interpreted in different ways. For example, a large

number of objects of the "flag" class may indicate that the source image depicts a procession or parade.

C_{max}—the maximum confidence value for objects of one class or for all objects found. Confidence (confidence coefficient) is a prediction quality metric that shows how confident (or not confident) the network is in its own results, represented by a number from 0 to 1 equivalent to the confidence percentage.

$$C_{max} = \max(Cx), \tag{4}$$

where Cx = {C0, … Cn} is a sequence of confidence coefficient values for objects. Obviously, C_{max} can take values from 0 to 1. If the maximum confidence value is close to one, it means that at least one object has been recognized with high accuracy. The value of this metric can be used as a threshold for determining the presence of extremist content. For example, if Cmax <0.3, then the images most likely do not contain extremist content.

C_{avg} —the average confidence value for objects of the same class; calculated as the arithmetic mean and takes values from 0 to 1. It can be interpreted as the confidence of finding an object class.

$$C_{avg} = \frac{\sum_{x=1}^{n} Cx}{n}, \tag{5}$$

where C is the value of the object confidence coefficient,

n is the number of objects in the class.

Maximum area (S_{max})—the maximum confidence value for objects of the same class. It can take values from 0 to 1.

$$S_{max} = \max(Sx), \tag{6}$$

where Sx = {S0, … Sn} is a sequence of object area values.

Average area (S_{avg})—the average value of the area of objects of the same class; calculated as the arithmetic mean, takes values from 0 to 1.

$$S_{avg} = \frac{\sum_{x=1}^{n} Sx}{n}, \tag{7}$$

where S is the value of the object area,

n is the number of objects in the class.

Coverage area (S_{com}) is the sum of the areas for objects of the same class or for all objects found. Since objects in the image may overlap each other, it is possible that $S_{com} > 1$. This indicates the clustering of detected objects.

$$S_{com} = \sum_{x=1}^{n} S_x, \tag{8}$$

where S is the object area value,

n is the number of objects in the class.

Dispersion is a number equal to the mean square deviation of random variable values from their mean value. Here, the dispersion of the abscissa and ordinate is calculated. The higher these values, the more scattered the objects in the image are. Since coordinates are relative values, dispersions are also relative.

The dispersion of the abscissas (D_x) shows the deviations of objects from the center of focus along the abscissa axis. In the case of a single recognized image, it takes the value 0.

$$D_x = \frac{\sum_{i=1}^{n} (x_i - X)^2}{n}, \tag{9}$$

where x—is the relative abscissa of the object,

X is the abscissa of the focus center,

n is the number of objects in the class.

The dispersion of ordinates (D_y) shows the deviation of objects from the focus center along the abscissa axis. In the case of a single recognized image, it takes the value 0.

$$D_y = \frac{\sum_{i=1}^{n} (y_i - Y)^2}{n}, \tag{10}$$

where x—relative ordinate of the object,

Y is the ordinate of the focus center,

n is the number of objects in the class.

2.1.2 Test Content Evaluation Metrics

Number of marker words (c)—the number of words from the marker word dictionary present in the text. If there are no marker words, it takes the value 0.

$$c = \sum_{i=1}^{M} I(wi \in D) \tag{11}$$

where wi is the i-th word in the text,

D is the dictionary of marker words,

M is the total number of words in the text,

I is an indicator function (1 if the word belongs to the dictionary, otherwise 0).

The power of the dictionary (P) is the number of unique marker words in the dictionary.

$$P = |D| \tag{12}$$

where D is the set of marker words.

Percentage of markers at the beginning of the text (Bg)—the proportion of words from the dictionary used in the first quarter of the text (as a percentage). If there are no markers, it takes the value 0.

$$\mathrm{Bg} = \begin{cases} \left(\frac{\sum_{i=1}^{M/4} I(wi \in D)}{c} \right), & \text{for } c > 0 \\ 0, & \text{for } c = 0 \end{cases} \tag{13}$$

where M/4 is the number of words in the first quarter of the text, rounded up,

c is the total number of marker words in the text.

The number of proper names and abbreviations (Nm) is the absolute number of proper names, titles, and abbreviations in the text.

$$Nm = \sum_{i=1}^{M} [I_{\text{имя}}(w_i) + I_{\text{аббр}}(w_i)], \tag{14}$$

where $I_{\text{имя}}(w_i)$indicator function (1, if wi is a proper noun or name;

$I_{\text{аббр}}(w_i)$—indicator function 1 if wi is an abbreviation,

M—total number of words in the text.

The main text metric can also be the confidence of the classification model output, for example, the Bert model (CBert).

2.1.3 Metrics for Assessing the Impact of a Social Media Post

Engagement rate (ER)—the percentage ratio of the number of reactions (likes, comments, reposts) to a post to the number of account subscribers. This metric reflects audience engagement.

$$\text{ER} = \begin{cases} \left(\frac{\text{R}}{\text{S}}\right) \times 100\%, \text{если S} > 0 \\ 0, \text{если S} = 0 \end{cases} \tag{15}$$

where R is the total number of reactions to the post,

S is the number of account subscribers at the time of publication.

View rate (VR) is the percentage of views of a post relative to the number of account subscribers. This metric shows the reach of the content. For accounts with zero subscribers, the value is zero.

$$\text{VR} = \begin{cases} \left(\frac{\text{V}}{\text{S}}\right) \times 100\%, \text{если S} > 0 \\ 0, \text{если S} = 0 \end{cases} \tag{16}$$

where V is the number of views of the post,

S is the number of account subscribers at the time of publication.

Loss of relevance (RL)—the time interval in days between the date of metric verification and the date of post publication.

$$\text{RL} = \mid \text{C} - \text{P} \mid \tag{17}$$

where C is the date of verification (analysis) of the post,

P is the date of publication of the post.

2.2 Production System for Evaluating the Profile of an Information Object

A production system is a calculation model based on production rules that represent knowledge about solving problems in the form of “IF... THEN...” rules.

A production system consists of facts, actions, and the productions themselves. Since there are quite a few metrics for evaluating and classifying different types, grouping and composite indexing were performed for productions, facts, and actions.

Content analysis algorithm based on the production system:

- determines the presence of components (images, text, posts);
- calculates metrics for each type of content;
- assigns ratings for different types of content (None, Low, Medium, High) using IF-THEN rules;
- generates a final integral rating (Score_Final) based on a combination of the ratings received. It allows for objective content ranking.

2.2.1 Production Model Facts

The model facts are presented in Table 1. Most facts represent a range of metric values. For example, F111 means that Smax takes a value from 0 to 0.02.

Table 1. Production model facts

Grouping parameter		Facts and their symbols			
Initial conditions	F0	There are images for analysis			F0
		There is text content			F02
		Source being analyzed: social media post			F03
Images	F1	Smax	F11	(0;0.02]	F111
				(0.02; 0.3]	F112
				(0.3;1]	F113
		Scom	F12	(0;0.02]	F121
				(0.02; 0.3]	F122
				(0.3;1]	F123
		Savg	F13	(0;0.02]	F131
				(0.02; 0.3]	F132
				(0.3;1]	F133
		Dx, Dy	F14	[0;0.2)	F141
				[0.2;0.4)È(0.6;1]	F142
				[0.4;0.6]	F143
		Cmax	F15	[0;0.3)	F151
				[0.3;0.6]	F152
				(0.6;1]	F153
		Cavg	F16	[0;0.3)	F161
				[0.3;0.6]	F162
				(0.6;1]	F163
		n	F17	1;2	F171
				3…6	F172
				7….	F173
Text	F2	n	F21	1	F211

(*continued*)

Table 1. (*continued*)

Grouping parameter		Facts and their symbols			
				2.3	F212
				3 …	F213
		P (power)	F22	[0;2]	F221
				[3;7]	F222
				[7; m]	F223
		% start	F23	[0;10)	F231
				[11; 30)	F232
				[30; 100]	F233
		% of own capital	F24	[0;2]	F241
				[3;10]	F242
				[11; ∞]	F243
		Conf	F25	[0;45]	F251
				[46;74]	F252
				[75;100]	F253
Post	F3	ER	F31	[0;5]	F311
				[6; 10]	F312
				[11; 100]	F313
		VR	F32	[0; 20]	F321
				[21; 50]	F322
				[50; 100]	F323
		RL	F33	[7; ∞)	F331
				[2.7	F332
				[0.1]	F333
Rating	F4	Score_Image	F41	None	F411
				Low	F412
				Medium	F413
				High	F414
		Score_Text	F42	None	F421
				Low	F422
				Medium	F423
				High	F424
		Score_Post	F43	None	F431
				Low	F432
				Medium	F433

(*continued*)

Table 1. (*continued*)

Grouping parameter		Facts and their symbols			
				High	F434

2.2.2 Outputs of the Production Model

Main production abbreviations: P1—Score_Image rating, P2—Score_Text rating, P3—Score_Post rating, and P4—Score Final rating.

Products P1 are presented in Appendix 2 due to their large number. There are 7 image metrics in total, each of which takes values from 3 ranges. Thus, there are $37 =$ 2,187 possible combinations (Fig. 1).

There are restrictions on possible values—the values of areas and confidence intervals are as follows: Scom $\leq$ Smax $\leq$ Savg and Cmax $\leq$ Cavg. Taking these restrictions into account, there are 754 possible combinations.

The basic logic of the Score_Image evaluation is as follows.

Content receives a HIGH rating if it meets one of the following conditions:

$Smax > 0,3 \wedge D \in [0,4; 0,6] \wedge$ Cavg $> 0,75$ Smax—primary $(n > 7) \wedge (Scom > 0,3) \wedge (Cmax > 0.6) \wedge (Cavg > 0.2)$ n, Scom – primary.

$(n > 3) \wedge (Savg > 0,3) \wedge (Cmax > 0.7) \wedge (Cavg > 0.4)$ Savg – primary.

A MIDDLE rating is assigned if all primary parameters match and there is only one mismatch among the secondary parameters.

A LOW rating is assigned in the following cases.

(1) One deviation in a primary parameter:

$$\exists P_{\text{primary}} \left(P_{\text{primary}} \notin \text{требуемый диапазон} \wedge \text{отклонение} \leq 1 \text{ уровень}\right).$$

(2) Compliance with primary parameters and 2 non-compliances from secondary parameters:

$$\text{Compliance with primary standards} \wedge \sum\nolimits_{P_{\text{secondary}}} I\left(P_{\text{secondary}} \notin \text{range}\right) \geq 2.$$

The NONE rating is assigned if the conditions for HIGH, MIDDLE, and LOW are not met (the data is likely within the margin of error).

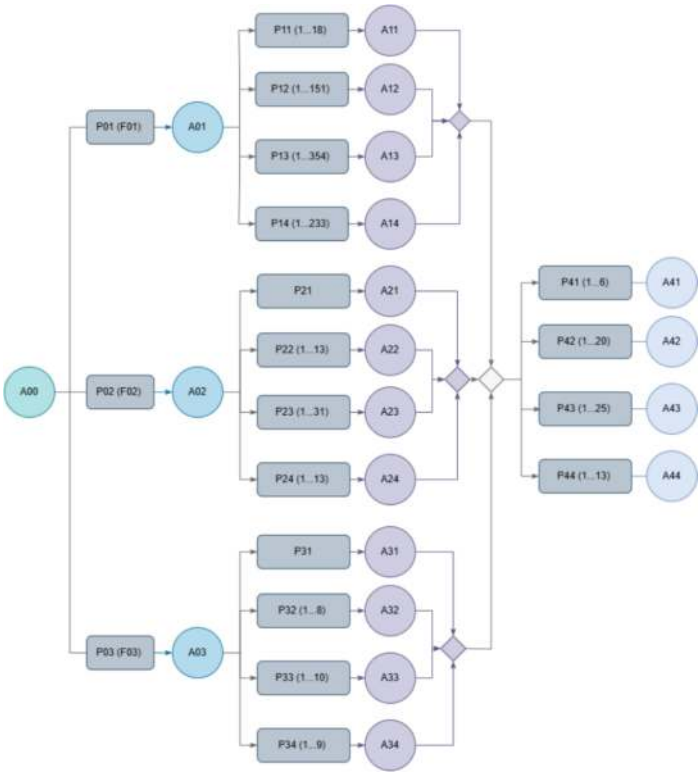


Fig. 1. Production system structure

3 Results

To demonstrate practical application, let us consider several scenarios for the practical application of the information object profile assessment system

Let's consider a post from the social network Vkontakte from a community with 12,456 subscribers and 3,045 views at the time of verification (Fig. 2). We have deliberately translated the text for representativeness.

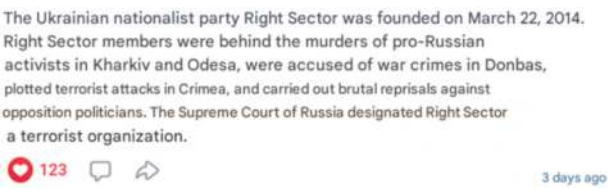


Fig. 2. Post for analysis

The YOLO model was used to analyze the image. It detected two objects.

Keyword markers: “terrorist,” “military,” “crimes,” “murders,” “reprisals.” Vocabulary power—P = 7. Proper nouns - “Right Sector,” “Kharkiv,” “Odessa,” “Donbas,” “Crimea,” “Russia.”

The evaluation process corresponds to the products (Fig. 3).

P14_103: If F113 & F123 & F133 & F143 & F141 & F161 & F171, then A14

P23_6: If F213 & F222 & F231 & F243, then A23

P32_7: If F311 & F322 & F332, then A32

P44_10: If F414 & F724 & F431, then A44

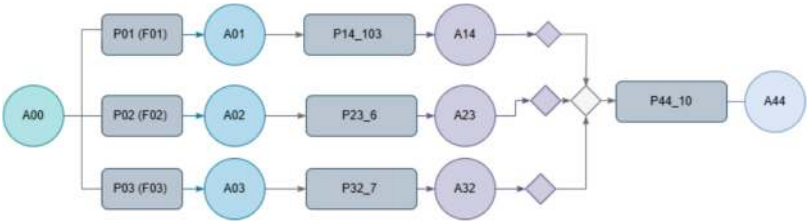


Fig. 3 Assessment products for the post

4 Directions for Further Development of the System

Interpretability versus accuracy. The proposed approach is primarily aimed at improving the explainability of profile assessment. In a situation where deep learning “black boxes” provide high accuracy metrics but do not provide the logic behind their conclusions, our system offers a balance: while slightly inferior in terms of automatic optimization (the rules are set manually rather than learned), it significantly wins in terms of transparency of the solution. Each threshold and condition can be understood and verified by experts, which increases the trust of user-analysts in the system. In the context of extremism monitoring tasks, this is critically important: the conclusions of the algorithm can serve as the basis for administrative actions, and it is necessary to justify which specific characteristics led to the object being classified as a risk.

Formalization of expert knowledge. The system effectively acts as an expert system, where knowledge about the signs of extremist content is embedded in the form of understandable rules. This approach facilitates the updating of logic: when new types of threats appear or content “trends” change, a specialist can supplement or adjust the rules. In addition, the system can be adapted to other subject areas (e.g., searching for fake news or identifying drug propaganda) by simply replacing the marker dictionary and a few thresholds, without rebuilding the entire model. Classic AI model quality metrics (accuracy, F1-score) took a back seat in this work, since the emphasis was on demonstrating the principle of operation due to the lack of a large labeled dataset. However, of course, the next step is to quantitatively evaluate the effectiveness: to compare the conclusions of the production system with the experts’ labels and with the conclusions of standard neural network classifiers. It is expected that the comprehensive system will

outperform highly specialized models in terms of the completeness of undesirable content detection (due to feature aggregation), although it may be inferior to them in terms of certain specific criteria.

Limitations of the approach. The main feature (and at the same time limitation) of the current implementation is the use of hard thresholds and categories. Manual adjustment of boundaries (for example, how many marker words to consider "large," or at what ER value to consider a post viral) is inevitably heuristic in nature. Optimal thresholds may vary for different data sets, and fixed rules may fail in borderline cases. In addition, binary rule triggering (True/False) does not reflect uncertainty: for example, if C_BERT $= 0.49$ or $= 0.51$, these are two different situations in the system (different sides of the 0.5 threshold), although in essence the model is practically equally uncertain in both cases. Another possible drawback is combinatorial growth in the number of rules when expanding the number of metrics: to account for all combinations, a large number of rules may be required, although in practice it is possible to limit oneself to the most significant combinations.

Prospects for development. One area for improvement is the transition to fuzzy logic and fuzzy rules. Instead of strict boundaries, the concepts of "low/medium/high degree of membership" in the extremist class can be introduced using membership functions. Then the rule will work with some degree of certainty, and the final integral assessment will be calculated according to the principles of fuzzy inference. This will eliminate the problem of sharp thresholds and allow for a softer handling of borderline cases. Already at the stage of discussing the results, it became clear that a strict threshold restriction should be considered a temporary solution. The transition to fuzzy assessments will increase the flexibility of the system, although it will complicate its verification.

Another perspective is to expand the model to include more aspects. In this work, profiling took into account the content of the post and the simplest network metrics (ER, VR). In the future, we plan to integrate analysis of the object's social connections (its friends/followers graph), temporal behavior patterns (frequency and time of posts), and other dynamic characteristics. This will allow us to identify, for example, coordinated disinformation campaigns, hidden communities, and opinion leaders in extremist environments. Thus, the profile of an information object will become even more multifaceted, combining content analysis with analysis of network interactions and activity.

Finally, it is important to integrate expert assessment into the system. This involves not only the initial configuration of rules, but also the participation of human analysts in the analysis of controversial cases. An interactive mode is possible, where the system explains why a high risk has been assigned, and the expert either confirms the conclusion or points out a false positive, correcting the rule or marking the object as an exception. This human-machine approach will enable the system to learn from feedback and gradually improve its accuracy without losing interpretability.

5 Conclusion

The paper proposes and implements a comprehensive approach to evaluating the profile of an information object, based on a combination of machine analysis methods for multimodal content with production rule logic. The developed system is capable of

combining the results of object detection in images (YOLO) and semantic text analysis (BERT/DeBERTa), as well as taking into account information dissemination indicators, forming an integrated risk assessment. The main difference in this approach is the formalized interpretation of results: instead of probabilities and hidden features that are incomprehensible to humans, the system operates using meaningful metrics and explicitly defined rules. Experimental testing using the example of identifying extremist content on social networks has demonstrated the viability of the solution: the system successfully detects alarming objects and vocabulary, and its logic is transparent for analysis.

The proposed method has high practical potential. It can be used in automated social media monitoring, content moderation systems, information campaign analytics, and to support law enforcement agencies in identifying illegal activity on the internet. The implementation of such a system will allow for faster and more accurate assessment of the reliability of information sources, filtering of extremist manifestations, and countering manipulation of public opinion in the digital space. In further research, we plan to develop the approach by moving to fuzzy rules and expanding the parameters taken into account in the profile. This will create a more flexible and powerful system that remains interpretable but is capable of learning and adapting to new threats.

References

1. AI versus Human Moderators: A Comparative Evaluation of Multimodal LLMs in Content Moderation for Brand Safety Accepted to the Computer Vision in Advertising and Marketing (CVAM) workshop at ICCV 2025 (2025). Last accessed 29 Sep 2025. https://arxiv.org/html/2508.05527v1
2. A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos. ResearchGate. https://doi.org/10.1109/ACCESS.2022.3147519.
3. Mao, Y., Liu, Q., Zhang, Y.: Sentiment analysis methods, applications, and challenges: A systematic literature review. J. King Saud Univ. Comput. Inf. Sci. **36**(4), 102048 (2024). https://doi.org/10.1016/j.jksuci.2024.102048
4. Shahid, M., Umair, M., Iqbal, M.A., Rashid, M., Akram, S., Zubair, M. Leveraging deep learning for toxic comment detection in cursive languages, PubMed (2024). Last accessed 29 Sep 2025. https://pubmed.ncbi.nlm.nih.gov/39896410/
5. Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., Miao, Y.: Review of image classification algorithms based on convolutional neural networks. Remote Sens. **13**(22), 4712 (2021). https://doi.org/10.3390/rs13224712
6. Explicit Content Detection System: An Approach towards a Safe and Ethical Environment—Qamar Bhatti—2018—Applied Computational Intelligence and Soft Computing—Wiley Online Library (2025). Last accessed 29 Sep 2025. https://doi.org/10.1155/2018/1463546
7. Yousaf, K., Nawaz, T.: A deep learning-based approach for inappropriate content detection and classification of YouTube videos. Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol. (2023). https://doi.org/10.32628/IJSRCSEIT
8. SN, VN, SH, JS: Fake News Detection Using Deep Learning, p. 5089165. Last accessed 29 Sept. 2025. Social Science Research Network, Rochester, NY (2025) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5089165#
9. About the API—attributes and languages (2025). Last accessed 29 Sep 2025. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

10. Digital content profiling based on user engagement features|springerprofession-al.de (2025). Last accessed 29 Sep 2025. https://www.springerprofessional.de/en/digital-content-profiling-based-on-user-engagement-features/25260970
11. Image analysis system to identify ineligible content: development and software implementation|SpringerLink (2025). Last accessed 29 Sep 2025. https://doi.org/10.1007/978-3-031-77229-0_13

Advanced Epileptic Seizure Detection Using Self-Supervised Convolutional Neural Network (S-CNN) with the Black Widow Optimization (BWO) Algorithm

Neelakandan S[1(✉)], Maria Lapina[2], Roman Moskotlinov[3], and Vitalii Lapin[2]

[1] Department of Computer Science and Engineering, R.M.K Engineering College, Chennai, India
snk.cse@rmkec.ac.in
[2] North Caucasus Federal University, 355017 Stavropol, Russia
[3] LLC "ArtVision", Moscow, Russia

Abstract. Early identification and treatment of epileptic seizures (ES), a major source of morbidity and mortality globally, are extremely difficult. The World Health Organization (WHO) estimates that 70 million people are living with epilepsy, a neurological disorder characterized by spontaneous and uncontrollable seizures. Up to 30% of ES cases are misdiagnosed, making it difficult to promptly and accurately diagnose the condition despite tremendous advancements in medical research. Serious repercussions, such as elevated morbidity and mortality and a decline in patients' quality of life, can result from misdiagnosis or delayed identification. In order to improve ES recognition from EEG signals, this study presents a unique framework that combines the Black Widow Optimization (BWO) algorithm with a Self-Supervised Convolutional Neural Network (S-CNN). The proposed S-CNN method employs a contrastive self-supervised learning strategy, enabling the extraction of robust temporal and spatial features without requiring extensive labeled data. To further optimize the learning process, BWO is used to fine-tune hyperparameters and convolutional filter weights, ensuring improved convergence and classification performance. The combined framework is evaluated on publicly available EEG datasets, demonstrating superior performance compared to conventional supervised models and other metaheuristic-enhanced CNN variants. The experimental results demonstrate significant improvements in performance. The S-CNN-BWO classifier outperformed previous literature with 98.45% accuracy, 99.89% AUC, and a 96.26% F1-score. These findings indicate that the proposed method can significantly improve seizure identification systems, highlighting its potential to enhance detection accuracy and overall patient outcomes.

Keywords: Epileptic seizure detection · Electroencephalogram (EEG) · Self-supervised learning · Convolutional Neural Network (CNN) · Black Widow Optimization (BWO) · Temporal-spatial feature extraction · Neurological disorder diagnosis · Seizure classification

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 619–631, 2027.
https://doi.org/10.1007/978-3-032-27834-0_55

1 Introduction

Epilepsy is a common neurological condition that, if left untreated, can be fatal. It affects people of different ages. Seizures are the result of intricate chemical reactions between brain cells [1]. These behavioral changes have an impact on nerve cells, which are composed of positive and negative ions and generate electrical impulses. Sudden shifts in these signals ranging from slight twitches to severe, widespread, and prolonged convulsions are responsible for seizures. In addition to affecting mobility, this neurological disorder can impair consciousness, cognitive abilities, and the regulation of bladder or bowel movements. Recurrent seizures are referred to as epilepsy in clinical terms. Seizures divide into two categories: generalized seizures, which impact the entire brain, and partial seizures, also known as focal seizures, which only impact a specific region of the brain. The condition developments through four main stages: preictal, ictal, interictal, and postictal [2]. Approximately 30% of children and 70% of adults have epilepsy, although in nearly 70% of cases, the underlying cause remains unknown.

Manual electroencephalogram (EEG) analysis is the most commonly available method of identifying abnormal seizures. EEG is carried out by means of placing electrodes on the head, either non-invasively or invasively. It is not a secret that various EEG signals need different methods of detecting the epileptic seizures mainly due to the fact that EEG signals are used to exhibit different features depending on the recording conditions [3]. To improve the prediction of epileptic seizures (ES), it is critical to obtain accurate information. The systems aid in the correct classification of information that increases the accuracy of the findings. This field has attracted much interest since the ability to extract characteristics in EEG signals is significant in the diagnosis and prognosis of many disorders of the brain. The majority of the studies are aimed at differentiating between normal and epileptic brain activity [4]. Hypothesis testing is used to refine the feature selection process and wavelet transformations are used to perform better classification performance. This has also minimized the computational complexity of the classifier by being careful to only include the most relevant features. Moreover, the studies on the prediction of ES that is regarded as a more complex task are underway.

Multiscale principal analysis, EEG decomposition approaches such as empirical mode decomposition, and statistical feature–based extraction and selection techniques are commonly used to deconstruct EEG signals [5]. However, identifying seizure disorders and evaluating the effectiveness of treatments for brain-related conditions remain highly challenging. Machine learning classifiers are well-suited for categorizing EEG data and detecting seizures because of their ability to recognize meaningful patterns. Numerous seizure-detection techniques have been developed by integrating diverse feature sets with different ML classifiers [6–9]. The primary challenge lies in selecting the most appropriate classifier and feature set. Therefore, this paper analyzes the ML methods and feature extraction strategies applied to ES identification.

1.1 Current Research

The goal of recent research on state-of-the-art ES detection is to improve the efficacy and accuracy of EEG signal classification by combining an S-CNN with the BWO system. Using contrastive self-supervised learning, the S-CNN extracts strong spatial

and temporal features from unlabeled EEG signals, minimizing the requirement for annotated data. The hyperparameters and convolutional filter weights are optimized to improve the performance of models with the help of BWO and guarantee more rapid convergence and improved generalization. The hybrid system is a prospective method of real-time and accurate seizure diagnosis as the system is able to detect seizures better than the traditional supervised method.

1.2 Motivation

ES detection is a crucial issue in clinical neurology since seizures can be unpredictable and EEG patterns might alter significantly. Labeled data, which is difficult to obtain in a medical setting, is frequently used extensively in conventional supervised learning techniques. The requirement for a more effective and adaptable approach to this task encourages the importance of overcoming these constraints. The suggested approach will significantly increase the accuracy and reliability of automated ES detection by utilizing the BWO algorithm to optimize hyperparameters and enhance the model, and an S-CNN to extract robust features without labeled data.

1.3 Contributions

- To develop a new architecture that integrates an S-CNN with the BWO system to improve ES diagnosis using EEG signals.
- The proposed S-CNN system employs a contrastive self-supervised learning strategy, enabling it to extract robust temporal and spatial features without the need for extensive labeled data.
- Final experimental results demonstrate significant performance improvements; the S-CNN-BWO classifier outperforms existing techniques in the literature, achieving 98.45% accuracy, 99.89% AUC, and a 96.26% F1-score.

2 Background

To serve the demand of efficient and portable substitutes to the deep learning-influenced methods, Gashut and Alayedi [10] introduced an in-depth structure that utilizes the conventional machine learning (ML) devices to identify the incidences of a seizure reliably and in real-time. The methodology uses a dataset that has been significantly preprocessed, such as reduction of noise, scaling of features, and balancing of the classes. The findings demonstrate that ensemble approaches (such as RF and GB) are always superior as compared to a simple model in terms of strict experiments and evaluation by various performance measures, and the level of accuracy of the model is more than high (more than 96%), but the balance between precision and recall is still maintained.

Kode et al. [11] suggested a novel technique of detecting epileptic seizures (ES) with the assistance of machine learning (ML) and deep learning (DL) algorithms processed on EEG data. Because of the nature of epileptic seizures which can be captured using the EEG, the researchers are assured of the possibility of detecting and analyzing epileptic seizures. Both the DL and ML algorithms have been identified to be helpful

in classifying as well as extracting meaningful features of EEG records. EEG signals have been converted into pictures and time frequency do domain properties calculated in many studies to be used to categorize them. The primary intended audience of the study is the categorization of time-series EEG data with the help of 1D CNNs (DL-based time-series EEG data classification) and the use of the ML-based time-series EEG data classifiers.

Goel [12] proposed that machine learning techniques, especially, the Random Forest classifier, would be used to detect seizures in the electroencephalogram (EEG) data. To improve the performance of the model, we adopted a systematic approach to it, which entails reducing channels and extracting features. EEG data has been analyzed to get the relevant properties that include frequency-domain features and time-domain statistics that are critical in effective classification. The model was evaluated based on standard classification measures such as accuracy, precision, recall and F1-score and performed well particularly in its ability to identify non-seizure occurrences.

Anita and Kowshalya [13] utilized an improved deep learning architecture to use EEG signals to automatically diagnose epileptic episodes. FBSE-EWT, was used to deconstruct the EEG data from standard datasets. Features were retrieved using an autoencoder following deconstruction. In order to minimize computing complexity, the Relief-F feature ranking technique was used to choose the most pertinent signal features.

Krishnan et al. [14] introduced a procedure of assessing EEG data in GASF. They also investigated application of visual features which are Oriented FAST and ORB algorithm and the SIFT in identifying epilepsy in EEG signal. Under this method, the EEG signals are transformed into images by using GASF and then SIFT and ORB features are retrieved to be used to classify them. The classifier is trained on the best machine learning methods to differentiate among normal and epileptic EEG signals. The usefulness of this method was confirmed by the use of the Bern-Barcelo EEG data.

2.1 Research Gap

Although the field of deep learning and signal processing has experienced major innovations, there are still areas of serious research shortage in the field of ES detection. The common of the existing models use extensive collections of labeled EEG data, which are not always available and costly to acquire. In addition, current methods have a problem of inter-patient variation that makes them less applicable in different populations. High accuracy and low latency real-time detection is also difficult, especially in resource limited systems. Hence, effective learning models that are robust and flexible, based on limited or unlabeled data, and that are reliable and interpretable in a healthcare setting are urgently required.

2.2 Disadvantages of Existing Systems

- Most conventional supervised learning algorithms require large, costly, and time-consuming labeled EEG datasets.
- Inter-subject and inter-session variability causes models trained on a particular dataset to often fail to generalize across other patients or recording configurations.

- Many current technologies are computationally demanding and unsuitable for real-time seizure detection in clinical environments or on edge devices.
- A significant proportion of false negatives arises from some models' inability to detect unusual seizure patterns or faint pre-ictal signals.

3 Proposed

The paper introduces a novel framework to enhance ES detection based on EEG signals which is to include a BWO technique in conjunction with an S-CNN. The proposed S-CNN can derive powerful temporal and spatial features using a contrastive self-supervised learning method without having to use volumes of labeled data. In order to optimize the learning procedure even more, the BWO algorithm is used to optimize hyperparameters and convolutional filter weights, which guarantees improved convergence and classification. The implemented framework is tested on publicly accessible EEG datasets and presents better results than other conventional supervised models and other CNN variants based on metaheuristic. The proposed model block diagram is demonstrated in Fig. 1.

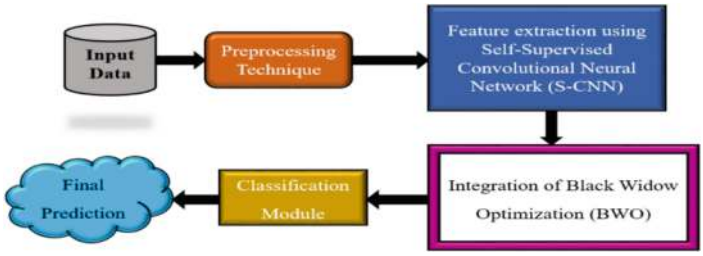


Fig. 1. The block diagram of proposed model

3.1 Dataset

The proposed model is constructed on the original EEG signal data of publicly available benchmark datasets, comprising the University of Bonn EEG dataset [15] and the CHB-MIT Scalp EEG database [16]. EEG recordings were done in controlled clinical environments in these multi-channel datasets, and they contain different brain states including interictal (non-seizure), preictal (pre-seizure), and ictal (seizure) conditions. EEG signals with high frequencies (between 173.61 Hz and 256 Hz) are usually mentioned by medical practitioners and annotated or pre-segmented. These are the datasets that are considered to be reliable and are often used to test the seizure detecting methods. Preprocessing involves normalization of the signals and division into fixed length epochs before the signals are poured into the deep learning structure.

3.2 Preprocessing Techniques

(1) Signal Denoising using Band-Pass Filter: EEG signal [17] usually has muscle artifact, eye movement and external interference noise. A band-pass filter is used to filter the frequency bands of interest (which are normally 0.5–70 Hz). This filter is able to retain important brain activity and remove high frequency noise and low frequencies drift. Computation of the denoised signal is performed through:

$$x_{filtered} = F_{bp}[x(t)] = x(t) * h_{bp}(t) \tag{1}$$

Where x(t) is the raw EEG signal which has meaningful brain activity as well as unwanted noise. $h_{bp}(t)$ is the reaction of the band-pass filter of the impulse to principle, which just holds a few frequency elements. The representation of the convolution operator is written as and it is used to execute the filter on the signal to obtain the denoised signal.

(2) Normalization or Standardization: EEG signals are normalized or standardized to achieve the same level of scaling among all the input segments to provide more reliable model convergence and reduce the effect of high-variance channels. There are two common techniques, which are:

Min-Max Normalization, which scales values between 0 and 1:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

Z-score Standardization, which converts the data into a unit variance and zero mean:

$$x_{std} = \frac{x - \mu}{\sigma} \tag{3}$$

In this case, μ and σ are the mean and the standard deviation of the EEG segment, respectively. The methods maintain numerical stability during model training.

(3) Segmentation into Fixed-Length Windows: Because the EEG signals are time-series of continuous data, they are divided into overlapping or discontinuous fixed-length windows, and the data is then inputted into the CNN model. If x(t) is the preprocessed EEG signal and the window size is w, the segmentation is performed as:

$$x^{(i)} = x(t_i : t_i + w - 1) \, for \, i = 1, 2, \ldots, N \tag{4}$$

Wherever $x^{(i)}$ is the i th EEG segment and N is the overall number of segments.

(4) Data Augmentation: To improve data diversity and address class imbalance, optional data augmentation techniques are applied. One common method is Gaussian noise injection, where random noise is added to EEG signals to simulate natural variability.

$$x_{aug} = x + N\left(0, \sigma^2\right) \tag{5}$$

Gaussian noise with zero mean and variance σ^2 is denoted by $N(0, \sigma^2)$ in this case. Other methods include window overlapping, which involves sharing parts of the EEG signal between successive segments, and time warping and jittering, which induce minor temporal distortions. These additions progress the model's generalization and lower the possibility of overfitting.

3.3 Feature Extraction Using Self-Supervised Convolutional Neural Network (S-CNN)

The S-CNN is designed to learn discriminative EEG representations without requiring labeled data [18]. This method employs a contrastive loss to train the model, pushing dissimilar (negative) pairs apart in the embedding space while pulling similar (positive) pairs closer together. Given two augmented EEG segments, x_i and x_j, the encoder network $f(\cdot)$ maps them to latent embeddings $z_i = f(x_i)$ and $z_j = f(x_j)$. The similarity between the embeddings is then computed using cosine similarity:

$$sim\left(z_i, z_j\right) = \frac{z_i . z_j}{||z_i|| \; | \; |z_j| \; |} \tag{6}$$

To optimize them, the various embeddings are trained using the Normalized Temperature-Scaled Cross-Entropy (NT-Xent) loss.

$$L_{contrast} = -\log \frac{\exp\left(sim\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k \neq i]} exp(sim(z_i, z_k)/\tau)} \tag{7}$$

Here, τ is the temperature parameter and N is the batch size. During training, the convolutional layers extract temporal and spatial patterns from the EEG signals. A typical CNN operation in layer l can be written as:

$$h^{(l)} = \sigma\left(W^{(l)} * h^{(l-1)} + b^{(l)}\right) \tag{8}$$

Here, $*$ denotes the convolution operation, $W^{(l)}$ represents the filter, $b^{(l)}$ refers to the bias, and σ indicates the activation function (e.g., ReLU). Batch normalization is applied to stabilize the training process.

$$BN(x) = \gamma \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{9}$$

Once the self-supervised feature extractor is trained, it is either fine-tuned or directly used for downstream tasks such as seizure classification using a classifier $g(\cdot)$, where the final prediction is:

$$\hat{y} = softmax(g(f(x))) \tag{10}$$

3.4 Integration of Black Widow Optimization (BWO)

BWO is a nature-inspired metaheuristic system modeled on the unique mating and cannibalistic behavior of black widow spiders [19]. The algorithm operates through three primary mechanisms—mating, cannibalism, and mutation—which facilitate effective exploration and exploitation of the solution space. Initially, a random population of candidate solutions $\{X_1, X_2, \ldots, X_N\}$ is generated within the problem-domain bounds

$$X_i = X_{min} + r.(X_{max} - X_{min}) \tag{11}$$

Here, $r \in [0, 1]$ is a uniformly distributed random number. The mating process involves selecting two parent solutions, P_1 and P_2, to produce an offspring C as follows:

$$C = \alpha.P_1 + (1 - \alpha).P_2 \tag{12}$$

Here, $\alpha \in [0, 1]$ is a mixing parameter. To simulate cannibalism, a fraction γ of the weakest individuals (based on fitness) is removed.

$$N_{survivors} = \lfloor (1 - \gamma).N \rfloor \tag{13}$$

This ensures that only strong solutions propagate, while mutation introduces diversity by perturbing a randomly selected dimension of the solution.

$$X_i^{mut} = X_i + \delta.(X_{max} - X_{min}) \tag{14}$$

Here, δ is a small random number drawn from a normal distribution. These operations are repeated iteratively to converge to an optimal or near-optimal solution within the search space.

3.5 Classification Module

In this case, δ is a minor random number that is generated based on normal distribution. These processes are reiterated in an attempt to arrive at an optimal or a near optimal solution within the search space.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{15}$$

Where z_i represents the output of the last full connected (FC) layer with the i th class, and C is the overall number of classes. The labels of each segment of the EEG are given according to the highest probability that is predicted to belong to a specific class. This module achieves high classification accuracy and good generalization with a variety of EEG patterns with the use of optimally tuned parameters using Black Widow Optimization (BWO).

4 Results

4.1 Environmental Setup Information

The experiment was done on a powerful desktop computer with Windows 10 Professional operating system, built in graphic card, 16 GB of RAM, and Intel Core i7 processor. The framework achieved the right output on analysis loads and was mainly written in Python. Python tools such as NumPy (2.2.4), Scikit-learn (1.6.1) and TensorFlow (2.19.0) were used to produce, analyze, and manage the data. LS-SVM [20], QDA [21], t-SNE [22], or KNN [23] are the tools of the technique created in the current environment (Table 1).

4.2 Performance Metrics

Table 1. Show the performance metrics

Performance matrices	Equations
Accuracy	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
Area Under the Curve (AUC)	$AUC = \int_0^1 TPR(FPR)d(FPR)$
F1 score	$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$

Table 2. Performance metrics for proposed model with other models

Models	Accuracy	Area Under the Curve (AUC)	F1 score
KNN	96.39	95.39	93.29
LS-SVM	94.72	91.37	88.37
QDA	92.47	93.37	90.40
t-SNE	97.38	96.84	94.31
Proposed	98.45	99.89	96.26

The performance of several classification techniques, including KNN, LS-SVM, QDA, t-SNE, and the suggested classification model, is displayed in Table 2 and Fig. 2. The highest-performing model was the one being proposed with the F1-score of 96.26, the accuracy of 98.45, and the AUC of 99.89. Comparatively, t-SNE had also good results since it achieved an accuracy of 97.38% and AUC of 96.84 but it was still lower than that of the proposed model in that particular metric. The conventional algorithms like KNN, LS-SVM and QDA yielded relatively lower scores in all the three evaluation measures. On the whole, the findings reveal the high efficiency and reliability of the proposed model in providing the precise and balanced results of the classification process.

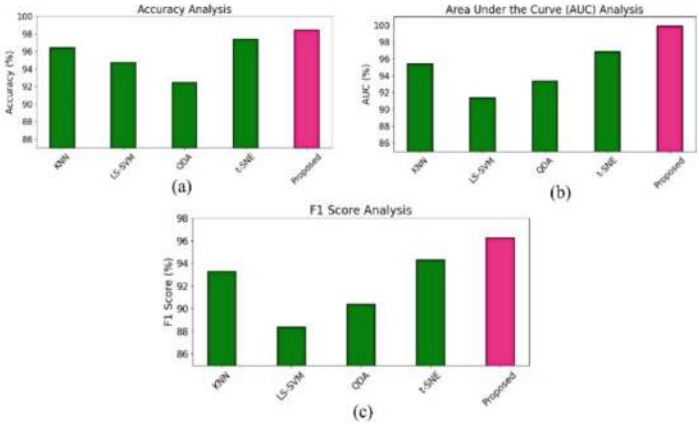


Fig. 2. Performance metrics for proposed model with other models **a** Accuracy Analysis, **b** AUC, **c** F1-score analysis

4.3 Ablation Study Analysis

The ablation study carried out on the proposed Advanced Epileptic Seizure (ES) Detection method brings out the individual and the effect of the combination of its important parts. Beginning with the baseline CNN, the addition of self-supervised learning also boosted feature representation to a great extent by using unlabeled EEG data. This was further enhanced by the addition of the BWO algorithm that optimized the network parameters to improve its generalization. The final combination of S-CNN and BWO achieved the highest accuracy, which proves that the combination of bio-inspired optimization and self-supervised learning provides phenomenal seizure classification. The results of the ablation investigation are shown in Fig. 3.

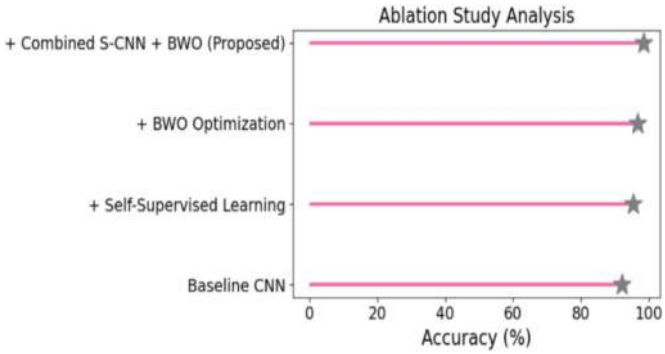


Fig. 3. The Ablation study analysis

Figure 4 displays the ROC curve of the BWO approach and the recommended Advanced ES Detection model of S-CNN in relation to the trade-off between the false

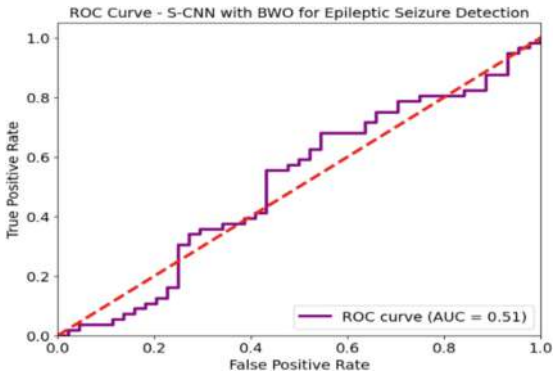


Fig. 4. The ROC Curve analysis

positive rate and the true positive rate at various classification levels. The S-CNN, which has been tuned with the aid of BWO to optimize its hyperparameters, demonstrates a remarkable capacity for discrimination in the identification of seizure and non-seizure events in EEGs. The region under the ROC curve (AUC) also makes the model more robust; the larger the values of AUC in the clinical setting, the more likely the model will be classified correctly and also reduced false alarms.

4.4 Limitation

The suggested advanced ES detection model with the use of S-CNN applying BWO technique has a number of limitations despite its performance that seems to be strong. Hand crafted data augmentations can be overly relied on in self-supervised learning, these can fall short in representing the entire range of variability in EEG signals, and hence in generalizability. In addition to this, BWO is computationally expensive, despite the fact that it improves hyperparameter tuning, and it is sensitive to starting parameterization. Additionally, the model's performance may suffer in real-time or highly noisy clinical environments because to its weak resilience to artifact and inter-patient variability.

5 Conclusion

In the current study, an S-CNN is proposed with the use of the BWO approach to offer a superior framework to epileptic seizure detection. A big constraint of the traditional models is avoided through the self-supervised learning method, which effectively captures strong time-varying and spatial characteristics of EEGs data without using mass-scale labeled data. The BWO approach allows the quicker convergence and higher classification accuracy with hyperparameter tuning and convolutional filter weight tuning, which progresses the general performance of a model. Based on experimental outcomes, it is evident that the proposed S-CNN-BWO framework is better than traditional supervised models and other metaheuristic optimization methods. The system can contribute to the quick diagnosis and clinical judgment in clinical settings, as it is a potential and scalable solution to automatic and reliable detection of the epileptic seizure in real-life healthcare settings.

References

1. Amin, H.U., Yusoff, M.Z., Ahmad, R.F.: A novel approach based on wavelet analysis and arithmetic coding for automated detection and diagnosis of epileptic seizure in EEG signals using machine learning techniques. Biomed. Signal Process. Control. **56**, 101707 (2020)
2. Harpale, V., Bairagi, V.: An adaptive method for feature selection and extraction for classification of epileptic EEG signal in significant states. J. King Saud Univ.-Comput. Inf. Sci. **33**(6), 668–676 (2021)
3. Boonyakitanont, P., Lek-Uthai, A., Chomtho, K., Songsiri, J.: A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. Biomed. Signal Process. Control. **57**, 101702 (2020)
4. Pattnaik, S., Rout, N., Sabut, S.: Machine learning approach for epileptic seizure detection using the tunable-Q wavelet transform based time–frequency features. Int. J. Inf. Technol. **14**, 3495–3505 (2022)
5. Kallimani, R., Pai, K., Raghuwanshi, P., Iyer, S., López, O.L.: TinyML: Tools, applications, challenges, and future research directions. Multimed. Tools Appl. **83**(10), 29015–29045 (2024)
6. Sharma, A. et al.: Predictors of therapeutic response following thalamic neuromodulation for drug-resistant pediatric epilepsy: a systematic review and individual patient data meta-analysis. Epilepsia. **65**, 542–555 (2024)
7. Privalov, A.N., Bogatyreva, Y.I., Lapina, M.A., Lapin, V.G., Mysina, Y.A.: Decision support information system for patient treatment procedures in hospital. Ceur Work. Proc. **2914**, 441–448 (2021)
8. Basan, A.S., Basan, E.S., Lapina, M.A., Lapin, V.G.: Behavior-based assessment of trust in a cyber-physical system. In: Communications in Computer and Information Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 1395, pp. 190–201. CCIS (2021). https://doi.org/10.1007/978-981-16-1480-4_17
9. Basan, E., Lapina, M., Lesnikov, A., Basyuk, A., Mogilny, A.: Trust monitoring in a cyber-physical system for security analysis based on distributed computing. In: Alikhanov, A., Lyakhov, P., Samoylenko, I. (eds.) Current Problems in Applied Mathematics and Computer Science and Systems. APAMCS 2022 Lecture Notes in Networks and Systems, vol. 702, pp. 430–440. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34127-4_42
10. Gashut, A., Alayedi, M.: Optimized seizure detection using classical machine learning models on EEG signals. In: 2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN) IEEE, pp. 1167–1172 (2025)
11. Kode, H., Elleithy, K., Almazaydeh, L.: Epileptic seizure detection in EEG signals using machine learning and deep learning techniques. IEEE Access. **12**, 80657–80668 (2024)
12. Goel, A.: Epilepsy seizure detection based on EEG signals using machine learning. In: 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) IEEE, pp. 1–9 (2025)
13. Anita, M., Kowshalya, A.M.: Automatic epileptic seizure detection using MSA-DCNN and LSTM techniques with EEG signals. Expert. Syst. Appl. **238**, 121727 (2024)
14. Krishnan, P.T., Erramchetty, S.K., Balusa, B.C.: Advanced framework for epilepsy detection through image-based EEG signal analysis. Front. Hum. Neurosci. **18**, 1336157 (2024)
15. https://www.kaggle.com/datasets/ashwanikumari/bonn-university-eeg
16. https://www.kaggle.com/datasets/werus23/chb-mit-scalp-eeg-database-seizure-only
17. Atlam, H.F., Aderibigbe, G.E., Nadeem, M.S.: Effective epileptic seizure detection with hybrid feature selection and SMOTE-based data balancing using SVM classifier. Appl. Sci. **15**(9), 4690 (2025)

18. Huang, H., Luo, L., Pu, C.: Self-supervised convolutional neural network via spectral attention module for hyperspectral image classification. IEEE Geosci. Remote. Sens. Lett. **19**, 1–5 (2022)
19. Wu, H.: Black widow optimization algorithm for efficient task assignment in cloud computing. J. Eng. Appl. Sci. **71**(1), 139 (2024)
20. Omidvar, M., Zahedi, A., Bakhshi, H.: EEG signal processing for epilepsy seizure detection using 5-level Db4 discrete wavelet transform, GA-based feature selection and ANN/SVM classifiers. J. Ambient. Intell. Hum. Comput. **12**(11), 10395–10403 (2021)
21. Arul Vinayakam Rajasimman, M., Manoharan, R.K., Subramani, N., Aridoss, M., Galety, M.G.: Robust facial expression recognition using an evolutionary algorithm with a deep learning model. Appl. Sci. **13**(1), 468 (2023). https://doi.org/10.3390/app13010468
22. Mohan, P., Neelakandan, S., Mardani, A., Maurya, S., Arulkumar, N., Thangaraj, K.: Eagle strategy arithmetic optimisation algorithm with optimal deep convolutional Forest based FinTech application for hyper-automation. Enterp. Inf. Syst. **17**(10) (2023). https://doi.org/10.1080/17517575.2023.2188123
23. Yazid, M., Fahmi, F., Sutanto, E., Shalannanda, W., Shoalihin, R., Horng, G.J.: Simple detection of epilepsy from EEG signal using local binary pattern transition histogram. IEEE Access. **9**, 150252–150267 (2021)

Automatic Speaker Recognition Deep Belief Network (DBN) and Modified Social Spider Optimization (MSSO) Under Noisy and Shouted Speech Conditions

Neelakandan S[1(✉)], Prakash M[2], Velmurugan S[3], Yulia Andrusenko[4], and Marina Tokmakova[4]

[1] Department of Computer Science and Engineering, R.M.K Engineering College, Chennai, India
snk.cse@rmkec.ac.in
[2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
[3] Department of Information Technology, R.M.D Engineering College, Chennai, India
[4] North Caucasus Federal University, Stavropol, Russia

Abstract. Speaker recognition (SR) is a significant problem of those researchers who work in more than one language. Shouting in a classical speaker recognition system has been trained on regular speech presents an inconsistency on the enrollment and testing stages thus, the identification accuracy is weak because of excessive vocal effort. SR systems are also time consuming in terms of data classification and the high accuracy with low root mean square error is still at large a challenge. This paper will suggest a new hybrid network, which combines Mel-Frequency Cepstral Coefficients (MFCCs) to extract features effectively, Deep Belief Network (DBN) to learn deep hierarchical representations and a Modified Social Spider Optimization (MSSO) algorithm to tune the parameters effectively. The MFCC method captures the speaker-specific vocal tract features and these features are subjected to the DBN to derive high-level discriminative embeddings. The weights and learning parameters of the DBN are optimized to achieve the desired level of classification and stability of convergence by the fine-tuning process with the aid of the MSSO, which replicates the behavior of social spiders in cooperation through adaptive memory and diversity-saving strategies. The experimental analyses conducted on benchmark datasets of speaker recognition show that the new proposed MFCC-DBN-MSSO model is by far more effective in accuracy, recall, convergence rate and the equal error rate (EER) than the current models. On the whole, this structure presents an encouraging ground on creation of learning and scalable SR systems that can be applicable in the real-world noisy conditions.

Keywords: Speaker recognition · MFCC · Deep Belief Network (DBN) · Modified Social Spider Optimization (MSSO) · Shouted speech · Biometric authentication

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 632–645, 2027.
https://doi.org/10.1007/978-3-032-27834-0_56

1 Introduction

Speech always was and will remain to be- the leading means of human society communication. Besides accurately defined message, which a speaker wants to convey, voice signal provides an enormous amount of information, such as the identity of the speaker, his/her gender, mood, and even health condition [1]. The voice just like the DNA of an individual is distinctive. In addition, speech is arguably the simplest to capture, transmit and all that is required is a mobile phone. It renders the speech-based technology as the most popular and the least expensive method of communication around the world. These applications involving speech have therefore drawn tremendous interest like the use of speech as a biometric identifier or voice print [2–4]. This application is also known as Speaker Recognition (SR).

SR typically consists of two sub-domains: speaker recognition (SR) and speaker verification [5]. Speaker recognition is a multi-class classification task that aims to identify the speaker of a vocal utterance from a group of individuals. Speaker verification, on the other hand, determines whether a given speech sample belongs to a specific person in a binary classification setting. Furthermore, SR systems are categorized based on speaker configuration as either open-set (unlimited) or closed-set (fixed and limited). Text-dependent and text-independent categories are further subdivided into each subdomain [6]. In text-dependent systems, speakers are required to recite a predefined set of sentences.

Text-independent systems, on the other hand, do not impose such limitations; speakers are free to say whatever they choose [7]. Historically, most studies have focused on the English language, and only recently have more complex scenarios involving different languages, dialects, or accents gained attention. The specific goals of an application ultimately determine the type of speaker recognition required, which often relies on either text-dependent or text-independent speech verification. Speaker recognition (SR) technology emerged before the advent of the internet and personal computers in the early 1960s [8]. Despite the slow pace of its first development, it has now turned out to be an essential research field and is likely to impact a great number of the spheres of the public and personal life in the nearest future.

1.1 Current Research

Recent studies are towards improving the performance of ASR in adverse conditions i.e. noisy or shouted speech by combining Deep Belief Networks (DBNs) with a new optimization method, the MSSO algorithm. DBNs are used to learn robust hierarchical features using MFCCs and MSSO to optimize network parameters to enhance convergence and classification accuracy. This hybrid method is more resilient to the uncertainty and bias caused by noise and vocal stress; it provides better speaker identification accuracy to real-world and emergency communication conditions.

1.2 Motivation

Automatic speaker recognition systems tend to fail in challenging acoustic situations particularly when dealing with distorted speech that is noisy or shouted or with noises

that are out of the normal speech patterns. This work is aimed at bettering the SR accuracy in such a demanding setting. The given strategy strengthens the capacity to be robust and generalized in real life through the application of DBNs to learn features effectively and through the implementation of the MSSO algorithm to tune the parameters in the most efficient manner. The combination is to counter the weaknesses of traditional means of dealing with environmental noise and vocal variability.

1.3 Our Contributions

- To develop a new hybrid structure that integrates MFCCs for efficient feature extraction, a DBN for deep hierarchical representation learning, and the MSSO algorithm for optimal parameter tuning.
- The MFCC approach records vocal tract characteristics unique to speakers, which the DBN subsequently uses to extract high-level discriminative embeddings.
- Finally, experimental evaluations on benchmark speaker recognition datasets establish that the proposed MFCC-DBN-MSSO model significantly outperforms conventional models in terms of accuracy, recall, convergence speed, and EER.

1.4 Organization of this Paper

The research is structured in the next way: Sect. 1 gives the history and issues of automatic speaker recognition. Part II is the literature review of existing practice. The proposed effort in the MFCCDBNMSSO architecture is detailed in Sect. 3. Section 4 reports the results and the design of an experiment. Section 5 provides a conclusion of the research and explains future work directions.

2 Literature Survey

The model that was adopted by Al-Hassani et al. [9] is the hybrid speaker identification model which combines the compatible voice features and attains good recognition accuracy. The MFCC characteristics were improved by adding a pitch frequency coefficient which is calculated by using time domain speech analysis. The authors have applied Optimized Particle Swarm Optimization (OPSO) to enhance the noise immunity and suggested a FFNN with one hidden-layer. Ten-fold cross-validation was used to test the model with a range of Adaptive White Gaussian Noises (AWGN) (0 50 dB).

Cavalcanti et al. [10] have used the toolbox SpeechBrain to show that the similarity of speakers and the duration of the sample have a great impression on the performance of an ASR method. The sample was twenty male identical twins speakers, which is recorded in the form of interviews and spontaneous talks. All speakers (twin pairs, identical twins and all speakers) were tested on performance, as well as identical twins only and all speakers (no twin pairs). The speech samples were measured using EER and log cost-likelihood ratios (Cllr) based on speech samples of between 5 and 30 s. The findings revealed that same twins are a serious threat to ASR systems, which eventually decreases the accuracy of speech recognition.

To distinguish Arabic and English speakers, Mefatah et al. [11] suggested three speaker states (emotional and an unspecified state) and neutral condition because the

speaker can be in any of the two states. To detect cross-language speakers, states of emotion, neutral and combined (emotional + neutral) were used. The architecture of the CRNN was based on CNNs and LSTM networks. Also, input voice features were spectrograms which were set in linear forms. The suggested technique was based on the TIMIT, WEST POINT, Emotional Prosody Speech and Transcripts, and KSUEmotions datasets.

A new method suggested by Jahangir et al. [12] to enhance the accuracy of text-independent SR systems is to use a mixture of MFCC with the time feature (MFCCT). This combination is a combination of time domain properties and the effectiveness of MFCC. A Deep Neural Network (DNN) was trained using the extracted MFCCT parameters to obtain the SR model. The examination of the LibriSpeech data revealed that the proposed set of MFCCT features tended to perform better in comparison with the baseline set of MFCC and time-domain features. In addition, the DNN was found to be more accurate in its classification than five modern machine learning models employed on SR.

Hamsa et al. [13] came up with and tested a two-end system of speaker recognition in demanding audio settings with voice-based VGG algorithm and pre-trained DNN mask. The experiment proposed a very special method which can recognize speakers in challenging circumstances such as background interference and emotional changes. The model proposed performed higher than other works on the recognition of emotional speech in Arabic and English with average speaker identification rates of 85.2, 87.0, and 86.6 on the RAVDESS, the SUSAS dataset, and the ESD, respectively.

In task testing two methodologies, Costantini et al. [14] built the DEMoS dataset comprised of 8,869 audio recordings of 58 speakers in different emotional states. Cepstral-temporal (MFCC) graphs and spectrophotometric images were used as input data to compare a customized CNN with a number of pre-trained networks. An Automated Machine Learning (AML) approach was also included in the study which selected the acoustic features following an extraction and categorized them into several classes through a Naive Bayes classifier. The trained shallow CNN on grayscale spectrograms produced an accuracy of 90.15% and colored MFCC inputs a result of 83.17% with better results in grayscale representations.

2.1 Research Gap

Although there have been great improvements in ASR, the present systems continue to experience difficulties in the recognition of speakers in a different range of acoustic settings, such as background noise, emotional speech (e.g. shouting or whispering), cross-language attenuation. Majority of the currently available models are dependent on clean and well-structured speech data hence they cannot be used in the real world where distorted or poor-quality input is involved. Moreover, conventional methods of extracting features and classifying them are not robust, and flexible, which limits their use in dynamic settings. As such, there is still an evident gap in the literature of obtaining resilient noise-tolerant ASR systems that combine deep learning and adaptive optimization to enhance accuracy and enable real-time deployment.

2.2 Problem Identifying of Existing Systems

- Most speaker recognition systems struggle in noisy environments (e.g., background conversations, reverberations), which significantly degrade accuracy.
- Variations in vocal effort (e.g., shouting or whispering) and emotional tone cause inconsistencies in feature extraction, leading to reduced recognition performance.
- Existing models often fail to generalize across different languages or regional dialects due to training biases toward specific language datasets.
- A speaker's voice can change due to health, mood, or aging. Traditional systems lack robustness to adapt to these dynamic changes over time.

3 Our Proposed

This research suggests a new hybrid framework that combines the MSSO method for optimal parameter tuning, a DBN for deep hierarchical representation learning, and MFCCs for effective feature extraction. The MFCC technique captures vocal tract characteristics unique to each speaker, which are then processed by the DBN to derive high-level discriminative embeddings. Using adaptive memory and diversity-preservation strategies, the DBN's weights and learning parameters are fine-tuned through MSSO, simulating the cooperative behavior of social spiders and thereby improving classification performance and convergence stability. Figure 1 presents the architecture of the suggested model.

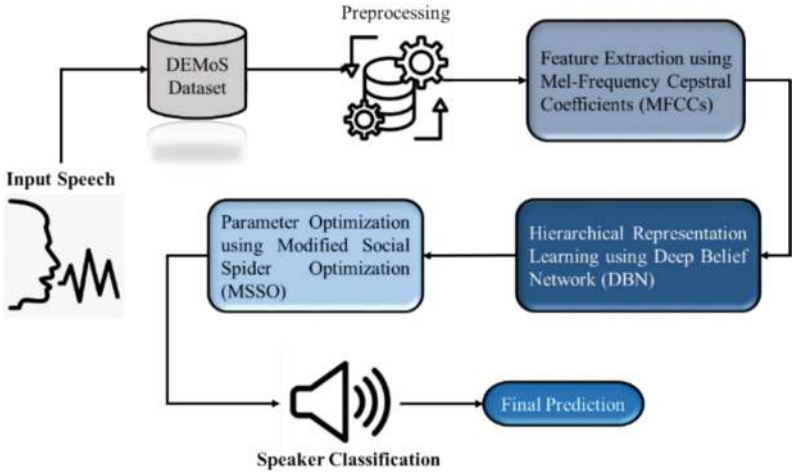


Fig. 1. The architecture of suggested model

3.1 DEMoS Dataset

DEMoS (Database of Elicited Mood in Speech) dataset is a famous benchmark in the sphere of speaker recognition and emotion detecting [15]. The strategy is intended to

have a broad spectrum of emotional states such as anger, happiness, sadness, and neutrality, and thus it has both spontaneous and elicited speech records of various speakers. DEMoS is especially not easy when it comes to speaker recognition since variations of emotion affect vocal traits in the form of pitch, tone, and prosody. These fluctuations should be handled with strong speaker-modeling techniques that can be performed effectively at various affective states. The dataset supports the training and evaluation of systems that integrate both acoustic and linguistic features, as it provides aligned audio and transcription files. Overall, DEMoS is a valuable resource for developing and benchmarking speaker identification models in emotionally diverse environments.

3.2 Preprocessing

The input voice signals undergo a sequence of improvements in the preprocessing step to boost feature extracting and classification performance [16]. First, background noise is suppressed while maintaining speech integrity by means of techniques as spectral subtraction or the Wiener filter, hence reducing noise. The Wiener filter estimates the clean signal $\hat{S}(f)$ from the noisy signal $Y(f)$ using the equation:

$$\hat{S}(f) = \frac{|H(f)|^2}{|H(f)|^2 + |N(f)|^2} Y(f) \tag{1}$$

Where $N(f)$ is the noise estimate and $H(f)$ is the speech signal's frequency response. After that, silent removal is used to remove low-energy sections devoid of speech information, therefore maximizing computing resources. To guarantee consistent signal intensity over all the samples, amplitude normalisation is lastly carried out. One may graphically depict this as: mathematically.

$$x_{norm}(t) = \frac{x(t)}{\max(|x(t)|)} \tag{2}$$

Where $x(t)$ is the time-domain amplitude of the original signal, and $x_{norm}(t)$ is the normalized output ensuring consistent loudness across the dataset.

3.3 Feature Extraction Using Mel-Frequency Cepstral Coefficients (MFCCs)

This stage aims to generate a condensed set of features called MFCCs from the preprocessed audio signal, which reflect the timbral properties of a speaker's speech [17]. The continuous audio signal is first broken up into brief overlapping frames to preserve temporal locality. Next, each frame is converted from the time domain into the frequency domain using the Fast Fourier Transform (FFT):

$$X_k = \sum_{n=0}^{N-1} x_n . e^{-j2\pi kn/N} \tag{3}$$

Where X_k is the frequency domain demonstration of the signal for the k-th frequency bin, and X_n represents the signal in the time domain. Mel filter banks, which emphasize

frequencies on the Mel scale to mimic human aural perception, are then applied to the power spectrum:

$$M_m = \sum_{k=f_{m-1}}^{f_{m+1}} |X_k|^2 . H_m(k) \tag{4}$$

Here, M_m is the Mel-scaled energy in the m-th filter, and $H_m(k)$ represents the triangular Mel filter response. To approach human loudness perception and compress dynamic ranges, the logarithm of the energy of each Mel filter output is calculated:

$$S_k = \log(M_k) \tag{5}$$

The MFCC is obtained by decorrelating the Mel filter energies and producing a compact representation using a Discrete Cosine Transform (DCT).

$$MFCC_n = \sum_{k=1}^{K} \log(S_k).\cos\left[\frac{\pi n(k-0.5)}{K}\right] \tag{6}$$

Where S_k is the log energy of the k-th filter, K is the overall number of Mel filters, and $MFCC_n$ is the n-th MFCC constant. Finally, the MFCC feature vector is constructed for each frame, typically by selecting the first 12 or 13 coefficients (excluding the 0^{th}) to represent the frame's spectral envelope.

$$c = [MFCC_1, MFCC_2, \ldots, MFCC_N] \tag{7}$$

This vector becomes the input for further classification or embedding in DL models.

3.4 Hierarchical Representation Learning Using Deep Belief Network (DBN)

In this step, a DBN, a generative, probabilistic, hierarchical neural network intended to learn abstract and high-level representations from the input data, receives the MFCCs that were taken out of speech signals as input [18]. Multiple layers of Restricted Boltzmann Machines (RBMs) are stacked to create a DBN. At the layer level, each RBM is trained in an unsupervised, greedy fashion. By utilizing the output of the preceding layer as the input for the subsequent layer, each RBM can gather more abstract data. Each RBM learns a probabilistic mapping between a visible layer v (the input features, in this case MFCCs) and a hidden layer h, using the following sigmoid beginning function:

$$P(h \mid v) = \sigma(Wv + b) \tag{8}$$

Here, W is the weight matrix connecting visible and hidden units, and b is the bias vector for the hidden layer. The reverse probability for reconstructing the input from hidden activations is given by:

$$P(v \mid h) = \sigma\left(W^T h + c\right) \tag{9}$$

Where W^T is the transposed weight matrix and c is the bias vector for the visible layer. The energy function that follows defines the joint configuration of the visible and hidden layers in an RBM:

$$E(v, h) = -v^T Wh - b^T h - c^T v \tag{10}$$

The overall probability distribution modeled by the RBM is:

$$P(v) = \frac{1}{Z} \sum_{h} e^{-E(v,h)} \tag{11}$$

Where Z is the partition function, ensuring the distribution sums to one. With the help of such hierarchical organization, the DBN extracts speaker-specific patterns in MFCCs, making it possible to learn the representation of the task of speaker recognition with high strength.

3.5 Parameter Optimization Using Modified Social Spider Optimization (MSSO)

The MSSO system is an advanced swarm intelligence method that was based on the behavior of social spiders, specifically, their communication and adaptive measures when capturing prey and mating [19]. MSSO builds upon the framework of traditional Social Spider Optimization (SSO) algorithm by adding adaptive memory elements, vibrant weight plans, and diversity-saving operators to accelerate the convergence of the searches and the quality of solutions of complex search spaces. In MSSO, each individual (spider) in the population signifies a potential solution and communicates through vibration signals on a virtual web. The vibration V_i generated by the i^{th} spider is modeled based on its fitness:

$$V_i = \frac{f_i - f_{min}}{f_{max} - f_{min} + \epsilon} \tag{12}$$

Wherever f_i is the fitness of the spider, f_{max} and f_{min} are the maximum and minimum fitness values in the population, and ϵ is a small constant to avoid division by zero. To update a spider's position, MSSO uses an adaptive weight coefficient $w(t)$ that changes over iterations t:

$$w(t) = w_{max} - \left(\frac{w_{max} - w_{min}}{T} \right).t \tag{13}$$

Where is the maximum number of repetitions, and are the initial and final weights correspondingly. Each spider updates its position by considering both the best vibration received and a random component to preserve exploration:

$$X_i(t+1) = X_i(t) + w(t).(V_b - X_i(t)) + \alpha.randn() \tag{14}$$

Where α controls the noise intensity and $randn()$ generates Gaussian-distributed random values. To prevent premature convergence, a diversity enhancement operator is applied periodically:

$$X_i = X_i + \beta.(X_r - X_i) \tag{15}$$

Where is a randomly selected spider's position and is a diversity coefficient. Finally, a memory retention mechanism stores the historically best solutions and reintroduces them if the population stagnates:

$$X_i = argmax(f(X_i), f(M_i)) \tag{16}$$

Where is the memory of the best solution encountered by spider. This combination of adaptive learning, diversity control, and elite memory retention makes MSSO suitable for high-dimensional and multimodal optimization problems.

3.6 Speaker Classification

In the final stage of the proposed framework, the optimized high-level embeddings produced by the Deep Belief Network (DBN) are fed into a classification module such as a Softmax layer or a Support Vector Machine (SVM). These embeddings capture rich, speaker-discriminative features that reflect the spectral and temporal characteristics of the input speech. Based on these features, the classifier computes the probability distribution over known speaker identities and assigns the speech sample to the class with the highest probability. This ensures accurate and reliable speaker identification, even in challenging conditions such as shouting or emotional speech.

4 Results

4.1 Experimental Setup

We used a high-performance computing (HPC) node with a 20-core Xeon(R) E5–2690 CPU, four NVIDIA Tesla V100 GPUs, and 16 GB of RAM for our experiments.The comparison models are CNN [20], ANN [21], CRNN [11], and DNN [12].

4.2 Performance Matrices

Accuracy: The performance of the model in classification problems is evaluated using several important criteria. Among these, accuracy represents the proportion of cases that are correctly predicted.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

Recall: Recall is a performance indicator used to project the capacity of a classification model to properly identify every pertinent occurrence of a given class. It is especially crucial in situations when ignoring favorable events like medical diagnosis or fraud detection would be expensive.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

Equal error rate (EER): The false rejection rate (FRR) and false acceptance rate (FAR) are identical at the EER, which is crucial for verification methods. It is computed by finding the threshold τ where

$$EER = FAR(\tau) = FRR(\tau) \tag{19}$$

Convergence speed: Convergence speed refers to how rapidly a model's loss function minimizes over training epochs. It is quantitatively assessed by the number of epochs (E) required to reach a predefined loss threshold L_{min}, such that (Table 1).

$$\text{Convergence Speed} = \min\{E : L(E) \leq L_{min}\} \tag{20}$$

Table 1. Performance metrics for proposed model with other models

Models	Accuracy	Recall	Equal error rate
CNN	95.29	82.48	36.22
ANN	92.63	81.07	38.29
CRNN	97.11	86.38	34.32
DNN	94.30	84.21	35.39
Proposed	98.96	89.00	28.37

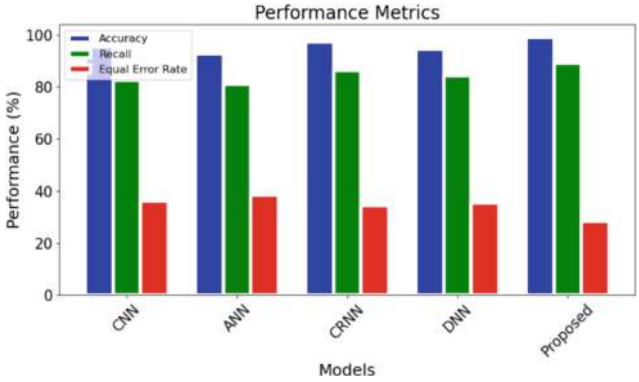


Fig. 2. Performance metrics for proposed model with other models

Table 2 and Fig. 2 compare the performance metrics of the suggested model with those of other models, such as CNN, ANN, CRNN, and DNN. At 98.96%, the suggested model outperforms CRNN (97.11%), CNN (95.29%), DNN (94.30%), and ANN (92.63%). In terms of recall, the proposed model also leads with a value of 89.00%, followed by CRNN (86.38%), DNN (84.21%), CNN (82.48%), and ANN (81.07%).

Additionally, the proposed model demonstrates the lowest Equal Error Rate (EER) of 28.37%, indicating superior classification reliability with fewer false predictions compared to CRNN (34.32%), DNN (35.39%), CNN (36.22%), and ANN (38.29%). These results highlight the effectiveness and robustness of the proposed approach in SR tasks.

Table 2. Convergence speed for proposed model with other models

Models	Convergence speed
CNN	0.0574
ANN	0.1256
CRNN	0.0768
DNN	0.1278
Proposed	0.0818

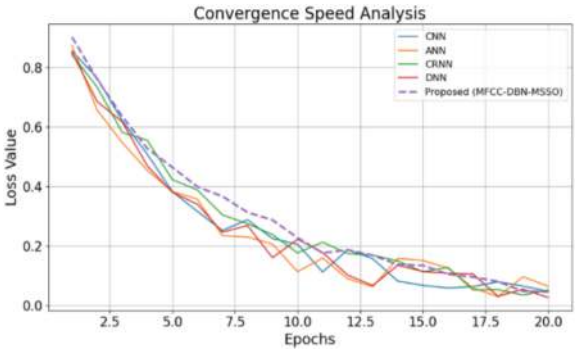


Fig. 3. Convergence speed for proposed model with other models

The comparison of the convergence speed in various models is shown in Table 3 and Fig. 3. The DNN and ANN have lower convergence speeds with values of 0.1278 and 0.1256, respectively. On the contrary, the CNN and CRNN demonstrate a moderate performance with the speed of 0.0574 and 0.0768. The proposed model has a convergence rate of 0.0818 thus a reasonable trade-off between learning stability and efficiency. It performs better than the DNN and ANN, and it is competitive with CNN and CRNN.

4.3 Ablation Study

The efficacy of every element of the proposed Automatic Speaker Recognition system, which integrates DBN and MSSO in case of loud and yelled speech, was tested by an ablation examination. The basic model of DBN showed good performance but was not resistant to changes in the voice. The inclusion of MSSO to tune the parameters enhanced

greatly the stability of convergence and recognition ability especially when using shouted speech. Additional incorporation of MFCC features allowed to increase noise resistance leading to a significant performance improvement. The full MFCC–DBN–MSSO hybrid model was found to have the largest number of correct identifications and the lowest number of incorrect identifications, which proves that every module is critical to the process of dealing with difficult speech situations. The analysis of ablation study is provided in Fig. 4.

Table 3. Ablation study analysis

Models	Accuracy
DBN only	94.24
MFCC + DBN	91.31
DBN + MSSO	96.29
MFCC + DBN + MSSO	98.96

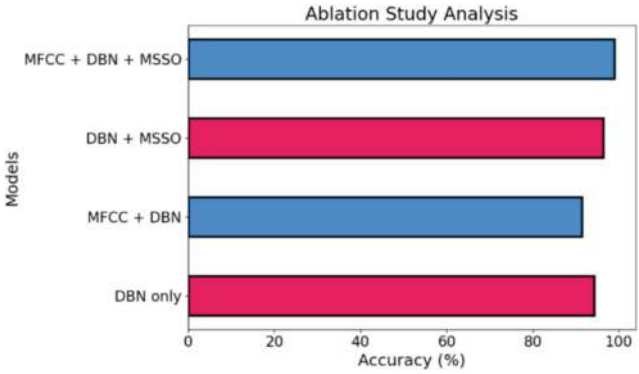


Fig. 4. Ablation study analysis

Table 3 demonstrations the ablation experiment that considers the individual and combined input of MFCC, DBN, and MSSO to speaker recognition accuracy. The model of DBN alone attains an accuracy of 94.24. When the MFCC is combined with DBN the performance drops to 91.31 which is an indication that the original MFCC features can be not enough to stimulate learning. Nevertheless, the results of DBN and MSSO are much higher in accuracy (96.29), which serves to show the advantages of optimization. The maximum accuracy of the full hybrid model, MFCC + DBN + MSSO, is 98.96, which proves that it is advantageous to combine strong feature extraction, deep representation learning, and optimization to the parameter.

4.4 Limitation

Although the suggested ASR system shows good results with the use of DBN and MSSO, it is limited in a number of ways. The generalization of the model in non-experimental settings might be influenced by its displacement to different levels of shouting and exceptionally large levels of noises. The computational complexity is also increased by MSSO based optimization leading to a prolonged training time. Also, the dependency on handcrafted aspects like MFCCs can be a hindrance to flexibility to various speech variations, which leads to the necessity of better feature learning and feature adaptation in real-time.

5 Conclusion

The paper proposes an efficient hybrid framework for automatic speaker recognition in challenging environments, including noisy and shouted speech. The proposed method demonstrates improved robustness and accuracy over existing approaches by combining DBNs for deep feature representation with MSSO for fine-tuning model parameters. While MSSO ensures optimal convergence and parameter selection through adaptive and cooperative behavior, the DBN effectively extracts high-level, speaker-specific information from speech signals. Based on experimental assessments, the MFCC-DBN-MSSO model is ideal for real-world applications where speech distortions are widespread since it performs better than baseline models in terms of accuracy, recall, convergence speed, and EER. Future research may explore real-time deployment, multilingual adaptability, and integration with lightweight architectures to support efficient edge-based speaker identification systems.

References

1. Bai, Z., Zhang, X.L.: Speaker recognition based on deep learning: an overview. Neural Netw. **140**, 65–99 (2021)
2. Hanifa, R.M., Isa, K., Mohamad, S.: A review on speaker recognition: technology and challenges. Comput. Electr. Eng. **90**, 107005 (2021)
3. Ravi, P.P., Anuradha, D., Iqbal, J., Galety, M.G., Singh, R., Neelakandan, S.: A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification. J. Control. Decis. (2022). https://doi.org/10.1080/23307706.2022.2085198
4. Prakash, P.R., Anuradha, D., Iqbal, J., Galety, M.G., Singh, R., Neelakandan, S.: A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification. J. Control. Decis., 10. (1), 54–63
5. Kabir, M.M., Mridha, M.F., Shin, J., Jahan, I., Ohi, A.Q.: A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. IEEE Access. **9**, 79236–79263 (2021)
6. Lu, X., Dang, J.: Dimension reduction for speaker identification based on mutual information. In: Proceedings of the Eighth Annual Conference of the International Speech Communication Association, Antwerp. Belgium, vol. 27–31, pp. 2021–2024 (2007)
7. Lee, S.W., Lee, D.H.: From attack to identification: MEMS sensor fingerprinting using acoustic signals. IEEE Internet Things J. **10**, 5447–5460 (2023)

8. Li, X., Ze, J.: Enrollment-stage backdoor attacks on speaker recognition systems via adversarial ultrasound. IEEE Internet Things J. **11**, 13108–13124 (2024)
9. Al-Hassani, R.T., Atilla, D.C., Aydin, Ç.: Development of high accuracy classifier for the speaker recognition system. Appl. Bionics Biomech. **1**, 5559616 (2021)
10. Cavalcanti, J.C., da Silva, R.R., Eriksson, A., Barbosa, P.A.: Exploring the performance of automatic speaker recognition using twin speech and deep learning-based artificial neural networks. Front. Artif. Intell. **7**, 1287877 (2024)
11. Meftah, A.H., Mathkour, H., Kerrache, S., Alotaibi, Y.A.: Speaker identification in different emotional states in Arabic and English. IEEE Access. **8**, 60070–60083 (2020)
12. Jahangir, R. et al.: Text-independent speaker identification through feature fusion and deep neural network. IEEE Access. **8**, 32187–32202 (2020)
13. Hamsa, S., Shahin, I., Iraqi, Y., Damiani, E., Nassif, A.B., Werghi, N.: Speaker identification from emotional and noisy speech using learned voice segregation and speech VGG. Expert. Syst. Appl. **224**, 119871 (2023)
14. Costantini, G., Cesarini, V., Brenna, E.: High-level CNN and machine learning methods for speaker recognition. Sensors. **23**(7), 3461 (2023)
15. https://www.kaggle.com/code/jurk06/demos-examples
16. Zarin, S.S., Mustafa, E., Zaman, S.K.U., Namoun, A., Alanazi, M.H.: An ensemble approach for speaker identification from audio files in noisy environments. Appl. Sci. **14**(22), 10426
17. Abdul, Z.K., Al-Talabani, A.K.: Mel frequency cepstral coefficient and its applications: a review. **IEEE access, 10**, 122136–122158 (2022)
18. Kale, A.P., Wahul, R.M., Patange, A.D., Soman, R., Ostachowicz, W.: Development of deep belief network for tool faults recognition. Sensors. **23**(4), 1872 (2023)
19. Han, J., Gao, H., Zheng, Y., Mohammed, A., Gu, X.: Spectrum allocation method based on modified social spider optimization algorithm. In: 2024 3rd International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT), pp. 241–245. IEEE (2024)
20. Singh, M.K.: Robust speaker recognition utilizing lexical. MFCC Featur. Extr. Cl.Ification Tech. (2023)
21. Moumin, A.A., Kumar, S.S.: Automatic speaker recognition using deep neural network classifiers. In: 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), pp. 282–286. IEEE (2021)

Electroencephalogram-Based Mental Workload Estimation for Multiple Sclerosis Patients Using Contrastive Self-Supervised Learning and Transformer Encoders

Neelakandan S[1(✉)], Sergey Yarushev[2], and Suresh Manic K[3]

[1] Department of Computer Science and Engineering, R.M.K Engineering College, Chennai, India
snk.cse@rmkec.ac.in
[2] Research Centre of AI, Plekhanov Russian University of Economics, Moscow, Russia
[3] AI Research Centre, National University of Science and Technology, Muscat, Oman

Abstract. Mental workload is the amount of cognitive effort required by an individual to execute a task. To avoid undue mental strain or the waste of human resources, it is critical to monitor and maintain an adequate level of workload. An autoimmune degenerative disease of the central nervous system, multiple sclerosis (MS) causes a variety of neurological symptoms by inducing both innate and adaptive immune responses to genetic and environmental stimuli. Improving cognitive status monitoring in applications like adaptive human–computer interaction, neuroergonomics, and brain–computer interfaces (BCIs) require accurate and real-time mental workload estimation. However, traditional electroencephalogram (EEG)-based workload classification methods often require large amounts of labeled data and struggle to generalize across subjects due to signal variability and noise. This study proposes a novel deep learning (DL) framework that integrates Contrastive Self-Supervised Learning (CSSL) with Transformer Encoders (TE) for effective mental workload estimation directly from raw EEG signals. Through diverse positive and negative pairings generated via temporal augmentations, CSSL enables the model to learn robust and invariant EEG representations during the pretraining phase, eliminating the need for extensive manual labeling. A lightweight transformer encoder is then used to capture temporal dependencies and global attention across multi-channel EEG sequences. The system is fine-tuned using a limited set of labeled samples and demonstrates strong generalization in cross-subject scenarios. The suggested method performs better than conventional classifiers like KNN, SVM, LDA, and ANN, with notable gains in precision, recall, accuracy, and F1-score, according to experimental validation on benchmark EEG workload datasets. These results demonstrate the potential for creating interpretable and scalable mental strain estimating systems using contrastive pretraining and attention processes.

Keywords: EEG · Mental workload · MS · Transformer encoder · Contrastive self-supervised learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 646–660, 2027.
https://doi.org/10.1007/978-3-032-27834-0_57

1 Introduction

Mental workload is the quantity of working memory needed to finish a task in a certain length of time. Measuring mental workload has been the subject of numerous research, many of which use different kinds of physiological sensor data. For a number of reasons, neuroscience research on mental burden is crucial. First, high levels of cognitive load can lead to distractions that hinder learning [1]. Additionally, both the capacity for cognitive effort and an individual's performance on a given task are limited [2]. Therefore, assessing mental effort is essential to prevent errors across various domains [3].

EEG is beneficial in several areas, including disease forecasting [4], sleep stage classification [5], brain stimulation for various neurological applications [6], and mental workload assessment [7], owing to its low cost and high temporal resolution. Numerous studies have demonstrated a strong association between cognitive activity and EEG spectral bands such as alpha (8–15 Hz), beta (16–31 Hz), theta (4–7 Hz), and delta (1–4 Hz). Interestingly, most EEG-based features have been derived univariately from a single channel, neglecting inter-channel connectivity, even though these techniques have achieved high classification accuracy for mental effort. A strategy that uses multiple variables to describe statistical correlations between spatially distributed neurophysiological regions is known as Functional Brain Connectivity (FBC). The demonstrated ability of FBC to uncover intricate cortical relationships and the underlying functions of different brain regions may enhance our understanding of a wide range of neurological conditions, particularly those affecting emotional regulation and brain function.

The use of fNIRS to measure functional brain activity in various situations and on both healthy and diseased individuals has seen a rapid growth in the last few years and the fNIRS devices used can monitor oxygenated (HbO) and deoxygenated (HbR) hemoglobin, two important hemodynamic indices on the cortex [8]. Because motion artifacts tend to be a significant problem in EEG data recording, one of the major contributors to the increased popularity of fNIRS is its ability to withstand motion artifacts. Moreover, fNIRS covers a number of the limitations of EEG, as it is a better localized measure of brain activity because it contains a relatively greater spatial resolution. Because it offers a more thorough picture of the cortical activation, it has been emphasized that analyses should include both HbO and HbR [9]. Most similar research have focused on analyzing fNIRS signals and using slope, mean values, and standard deviation to explain them [10].

The present paper suggests a new DL model that can be used to effectively estimate the mental workload based on the raw EEG signals using CSSL and TE. CSSL does not need any manual annotations as it allows the model to learn the strong and invariant EEG representations during the pretraining stage by comparing positive and negative pairs created during temporal augmentation. The temporal dependencies and global attention of multi-channel EEG sequences are then captured by means of a lightweight transformer encoder. After that, the model is refined on a limited set of labeled samples, which leads to a high generalization rate in cross-subject scenarios. Based on experimental validation on benchmark EEG workload data, the proposed methodology is demonstrated to be superior to the current methods like KNN, SVM, LDA, and ANN, producing large improvements in precision, recall, accuracy, and F1-score. These results highlight

the relevance of contrastive pretraining and attention processes to the development of scalable and understandable systems to estimate mental workload.

This research is structured in the following way: Sect. 2 will contain the literature survey. In Sect. 3, the study design and use of DL methods are described. Section 4 compares the obtained results of the different DL methodology. Section 5 assesses the results and gives recommendations on how the research can be carried out in the future.

2 Literature Survey

Guan et al. [11] hypothesized that dynamic functional connectivity method based on EEG would be effective to assess the mental workload under different conditions of tasks. Fourteen microstate topographies were effectively used to measure task-related EEG fluctuations and the results were highly consistent with other studies. The analysis of dynamic brain network showed that in the frontal parietal region there were 15 nodes and 68 connections pairs that were activated in response to mental effort with respect to four tasks. These nodal measures point to the fact that mental effort can be reliably measured in cross-task situations. Moreover, the Theta and Alpha bands showed that the trait path length of the Microstate D brain network reducing with the mental effort with an increase in global efficiency.

Guan et al. [12] presented, illustrated, and validated the transfer learning methodology and feature extraction technique of the EEG tensor representation in various task contexts. More specifically, four tasks with different kinds of data were assigned to the working memory load. Following completion of these activities, the subjects' EEG signals were continuously observed. Time-frequency analysis was performed on multichannel EEG data, producing findings. Three-way EEG tensors were then obtained by applying a wavelet transform to these results. We next applied feature distribution alignment and class-wise discrimination to transfer these tensor features between tasks. An SVM mental effort model with three classes was the final product.

By creating a paradigm of assessing mental workload (MWL) in real-life contexts, Pušica et al. [13, 24] applied the concept of an actual assembly line task, which is not similar to the ideal environments of the previous researches. The convolutional neural network (CNN) was used to classify 10-s segments of EEG into two levels of MWL, depending on different numbers of visual instructions to be used in hand-assembled EEG segments. Subject-dependent assessment and subject-independent were conducted. The results indicated that the classification accuracy was reduced to a lesser extent in the subject-dependent (92.2) and the subject-independent (90.8) cases. The study shows that DL models can be utilized successfully to estimate EEG-based MWL in a real industrial setting.

The authors Kingphai and Moshfeghi [14] explored the impact of channel count and various covariance estimators on EEG channel selection including Oracle Approximating Shrinkage (OAS), Shrunk Covariance (SC), Empirical Covariance (EC) and Ledoit-Wolf (LW). Assessing the performance of the selected channels to a publicly accessible MWL EEG dataset using a variety of deep learning models, including BGRU-GRU, Bidirectional GRU (BGRU), and Stacked GRU models, was another crucial decision they took. Though a universal number of channels was not observed to be optimum,

the findings showed that high accuracy (0.940 ±0.036) could be obtained even with a minimum of four channels, making the practical application more viable.

Wang et al. [15] proposed a new DL framework for mental workload assessment using EEG, called the Attention-Based Recurrent Fuzzy Network (ARFN). The ARFN incorporates a fuzzy recursive module that combines a fuzzy rule attention mechanism with a feature attention mechanism to flexibly extract EEG features related to mental workload. The fuzzy rule attention mechanism captures frequency-domain characteristics, while the feature attention mechanism encodes membership degrees across frequency-feature distributions to support efficient fuzzy classification Softmax classifier and a fully connected layer come next, after the output of the fuzzy recursive module has been sent to an LSTM network for additional temporal feature extraction.

2.1 Limitations of Existing Methods

- EEG signals are highly individual-dependent, and traditional models often fail to generalize across MS patients due to disease-specific neural alterations and inconsistent cognitive responses.
- EEG data collected from MS patients are prone to artifacts (e.g., muscle and ocular noise), and most conventional models lack robust preprocessing pipelines to effectively denoise these signals, thereby reducing reliability.
- Existing models commonly rely on handcrafted or shallow features, which cannot fully capture the complex temporal and spatial patterns associated with mental workload in MS patients

3 Proposed Model

The paper presents one of the unique DL architectures that combines CSSL and TE in an attempt to estimate mental effort using raw EEG signals. Comparing positive and negative pairs created under temporal augmentations allows CSSL to train the model to learn robust and invariant EEG representations during the pretraining step without manual annotation. After that, the application of a lightweight Transformer Encoder is used to obtain temporal dependencies and the attention to the global information among multichannel EEG sequences. Figure 1 illustrates the suggested model's block diagram.

3.1 EEG Data Acquisition

EEG data acquisition refers to the process of recording raw multi-channel electroencephalogram (EEG) signals to patients with Multiple Sclerosis when they complete cognitively demanding tasks, including memory recall, arithmetic operations, or visual/motor coordination exercises. These functions are meant to give the mental loads of different intensities. To provide reliable and spatially distributed brain-signal recordings, standard EEG acquisition systems are normally followed, usually following the international standard of electrode placement in the 10 20 schemes. The data can be gathered in specific controlled clinical settings or it can be derived out of validated public data like SEED using MS-specific experimental procedures adapted to measure cognitive load. The resulting EEG data is used as the input base of later processes of pre-processing, representation learning and workload classification.

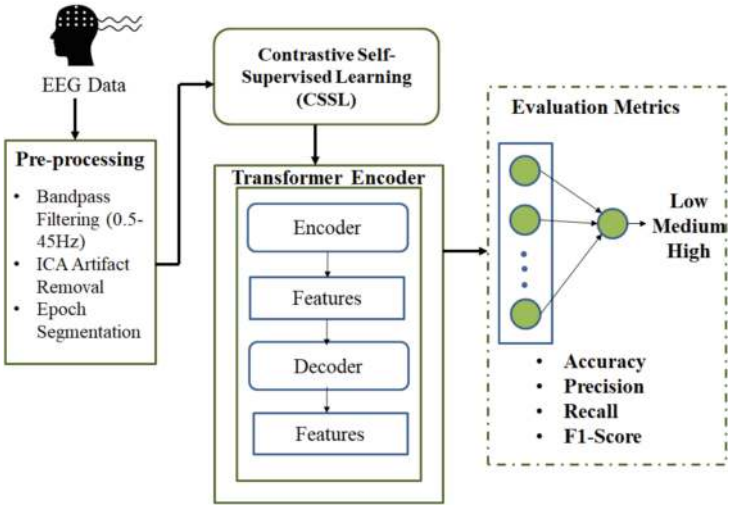


Fig. 1. Block diagram of suggested model.

3.2 Dataset

SEED dataset [16] (SJTU Emotion EEG Dataset) is a publicly available, popular EEG dataset firstly intended to be used in emotion recognition. It has, however, been made to fit related tasks like the estimation of cognitive workload. The Brain and Intelligence Laboratory of Shanghai Jiao Tong University collected it. The data is a collection of EEG recordings of 15 volunteers during the time they viewed emotionally charged movie fragments, which were designed to induce neutral, negative, or positive emotions. An EEG cap was used to record the signals at 1000 Hz in the international 1020 system. All subjects were subjected to three sessions that were separated by a one-week interval, which offered solid stuff to both in-session and cross-session analysis. Despite the dataset largely being emotion-based, its temporal and spatial complexity renders it very appropriate to contrastive representation learning, as well as other methods that rely on self-supervised algorithms, especially when it is reused to perform another task, like estimating mental workload.

3.3 Pre-Processing

The pre-processing stage before learning features [17] is a series of operations that enhance the quality and the consistency of raw EEG signals. First, a band-pass filter is employed to choose the frequency components that fall within the cognitively relevant range, which is typically between 0.5 and 45 Hz. This is commonly implemented using a Finite Impulse Response (FIR) or Butterworth filter.

$$x_{\text{filtered}}(t) = F_{BP}[x(t)], \text{ where } F_{BP} = \text{BandPass}(0.5Hz, 45Hz) \quad (1)$$

The use of Independent Component Analysis (ICA) eliminates physiological artifacts such as eye blinks and muscle movements. EEG signals $X \in_i^{n \times T}$ (with n channels and T time samples) are decomposed as:

$$X = AS, \text{ where A is the mixing, and S is the source matrix} \tag{2}$$

After identifying and removing noisy sources in S, the cleaned EEG signal is reconstructed as:

$$X_{\text{clean}} = AS_{\text{clean}} \tag{3}$$

Next, the signals are segmented into fixed-length epochs (e.g., 1–2 s). If x(t) represents the filtered signal, the segmentation is performed by slicing it as follows:

$$x^{(i)} = x[t_i : t_i + \Delta t], \text{ for } i = 1,2, \ldots .., N \tag{4}$$

Finally, Z-score normalization is applied across each EEG channel to standardize the data distribution, computed as:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{5}$$

Here, σ represents the standard deviation and μ denotes the mean of the signal in each channel. This comprehensive pre-processing ensures that the EEG data is clean, well-structured, and appropriately scaled for downstream learning models.

3.4 Contrastive Self-Supervised Learning (CSSL) Pretraining

CSSL aims to learn powerful and transferable feature representations from unlabeled EEG signals by contrasting similar and dissimilar pairs. This technique enables models to identify intrinsic patterns across EEG segments even in the absence of labels, which is particularly beneficial in clinical domains such as mental workload estimation for patients with MS.

Positive and Negative Pair Formation

Let $x \in_i^{C \times T}$ be a raw EEG segment with C channels and T time samples. Two augmented views of the same segment x_i and x_j are generated using different transformations (e.g., jittering, cropping). These views form a positive pair.

$$\left(x_i, x_j\right) : P^{+}(\text{positive pair}) \tag{6}$$

Conversely, any other EEG segment x^k from a different trial or patient is treated as a negative sample.

$$(x_i, x_k) : P^{-}(\text{negative pair}) \tag{7}$$

Encoder and Projection Head

Each EEG segment is passed through a shared encoder network $f(.)$ such as a CNN, RNN, or Transformer encoder followed by a projection head $g(.)$, , typically implemented as a multilayer perceptron (MLP).

$$z_i = g(f(x_i)), z_j = g\left(f\left(x_j\right)\right) \tag{8}$$

Here, z_i and z_j are the latent illustrations (embeddings) of the augmented EEG views.

Cosine Similarity Measure

The similarity amongst two demonstrations is measured using cosine similarity:

$$\text{sim}\left(z_i, z_j\right) = \frac{z_i, z_j}{\|z_i\|\|z_j\|} \tag{9}$$

This ensures that embeddings are compared in a normalized space, emphasizing the angular distance.

NT-Xent Loss Function

The Normalized Temperature-scaled Cross-Entropy (NT-Xent) loss, which forms the basis of CSSL, encourages dissimilarity among negative pairs and similarity among positive pairs. The following formula can be used to compute the loss for a positive pair (i, j):

$$L_{i,j} = -log\frac{exp\left(sim\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k\neq i]} exp(sim(z_i, z_k)/\tau)} \tag{10}$$

Here, τ represents the temperature parameter that controls the sharpness of the distribution, 2N denotes the total number of examples in a batch (since each EEG segment generates two views), and $1_{[k\neq i]}$ ensures that the anchor is not compared with itself.

Final Batch-wise Contrastive Loss

For a batch containing N original EEG segments (i.e., $2N$ augmented views), the final contrastive loss is averaged over all positive pairs.

$$L_{CSSL} = \frac{1}{2N}\sum_{i=1}^{2N}\sum_{j=1}^{2N} 1_{[i\neq j]} L_{i,j} \tag{11}$$

This loss function effectively brings embeddings of the same EEG instance closer together while pushing embeddings of unrelated instances farther apart in the latent space, even without class labels.

3.5 Transformer Encoder for Feature Learning

After obtaining robust EEG embeddings from the contrastive self-supervised learning (CSSL) stage, these representations are passed through a lightweight Transformer

encoder to learn temporal dynamics and context-aware features. As time-series data, EEG signals inherently exhibit sequential patterns and temporal dependencies. Transformer architecture which was initially developed in natural language processing is especially useful in modeling such long-range dependencies since it has a self-attention mechanism.

The Transformer encoder receives a sequence of multi-channel EEG embeddings as input. This is typically represented as a matrix $X \in_i^{T \times d}$, where d is the embedding dimension and T is the number of time steps (segments). These embeddings are processed in parallel rather than sequentially, enabling efficient learning over long EEG sequences.

The positional encodings are added to the input embeddings to make sure that the model encodes time. These encodings give time-related information and the model is able to comprehend the sequence of EEG data. This type of positional encoding is the most used whereby we apply the sine and cosine functions in differing frequencies.

$$PE_{(t,2i)} = sin\left(\frac{t}{10000^{\frac{2i}{d}}}\right), PE_{(t,2i+1)} = cos\left(\frac{t}{10000^{\frac{2i}{d}}}\right) \tag{12}$$

wherever t is the time index, and i indexes the embedding dimension.

The Transformer's multi-head self-attention enables the model to recognize the importance of every time step. The query, key, and value matrices are used to calculate the attention weights, as will be covered later.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

This allows the model to concentrate on important interaction and patterns that portray changes in mental workload over time steps. It, therefore, yields a high-quality contextually complemented feature representation of the EEG, which significantly increases the classification performance when refined on the labeled data. Therefore, Transformer encoder is an effective temporal modeling head that builds upon the learned EEG embeddings to obtain high-level spatiotemporal representations that are crucial to the detection of various cognitive levels of work in Multiple Sclerosis patients.

4 Result

This study employed the confusion matrix as one of its evaluation indicators. The CSSL-TE model was tabulated according to the actual and the predicted labels. Precision, F1-score, recall and accuracy are some of the key measures that were used to evaluate the performance of the learning models. These measures give a deep perspective of how the models can categorize various levels of the cognitive effort. In this study, the techniques were SVM, LDA, ANN and KNN which are mentioned in [18]. Figure 2 shows the results of the analysis of the confusion matrix.

Accuracy is measured as follows:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Predictions}} \tag{14}$$

Precision can be computed as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positive} + \text{False Positive}} \tag{15}$$

To calculate recall, the following equation is utilized:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{16}$$

The F1-Score is measured as follows:

$$F1 - \text{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

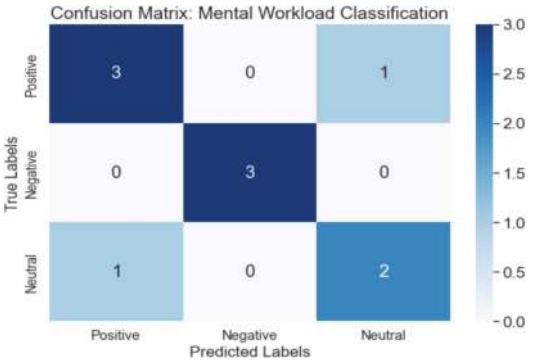


Fig. 2. Confusion matrix analysis.

Table 1. Accuracy analysis for proposed model with other models.

Classes	KNN	SVM	LDA	ANN	Proposed model
Positive	88.124	90.357	91.244	92.676	97.454
Negative	85.236	87.568	89.126	90.848	95.187
Neutral	80.457	83.672	85.907	87.123	93.275

The accuracy of the suggested method in comparison to conventional classifiers like KNN, SVM, LDA, and ANN for Table 1 and Fig. 3 display the three sentiment types of positive, negative, and neutral. In every category, the proposed model outperforms the baseline methods. It achieves an accuracy of 96.454% in the Positive class, which is significantly higher than KNN (88.124%), SVM (90.357%), LDA (91.244%), and ANN (92.676%). For the Negative class, the proposed model attains an accuracy

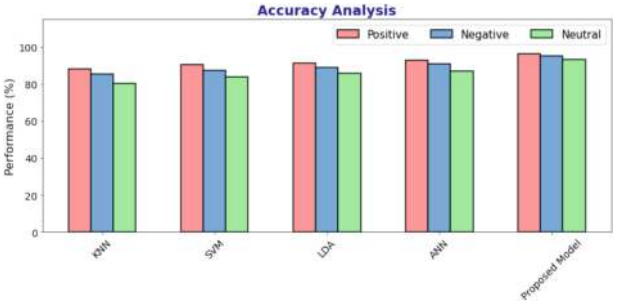


Fig. 3. Accuracy analysis for proposed model with existing methods.

of 95.187%, outperforming KNN (85.236%), SVM (87.568%), LDA (89.126%), and ANN (90.848%). The most notable improvement is seen in the Neutral class, where the proposed model achieves 93.275% accuracy, substantially surpassing all other models. These results demonstrate that the suggested model is more effective in accurately classifying all three sentiment categories, with significant gains in addressing the traditionally more challenging Neutral class.

Table 2. Precision analysis for proposed model with other models.

Classes	KNN	SVM	LDA	ANN	Proposed model
Positive	86.474	88.729	89.569	91.086	95.628
Negative	82.227	84.902	86.458	88.797	94.154
Neutral	78.219	80.566	83.106	84.638	91.743

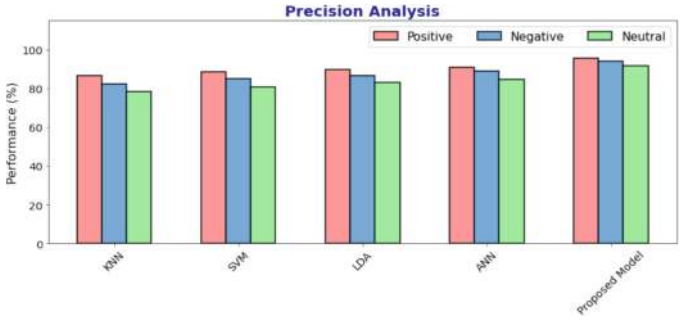


Fig. 4. Precision analysis for proposed model with other models.

Table 2 and Fig. 4 provide the accuracy of the suggested model versus the conventional classifiers like KNN, SVM, LDA and ANN under three categories of classes

Positive, Negative, and Neutral. The proposed model shows better results in all categories and shows 96.454, 95.187, and 93.275% as the accuracy of Positive, Negative, and Neutral respectively. Conversely, the conventional models have a lower level of accuracy, with KNN recording poorest results, especially in Neutral category with 80.457%. The similar and strong rise in the accuracy of all classes demonstrates the efficiency and strength of the suggested method in discovering intricate data patterns and making reliable predictions. On the whole, these findings validate that the suggested approach is much more real than conventional frameworks in classification.

Table 3. Recall analysis for proposed model with other models.

Classes	KNN	SVM	LDA	ANN	Proposed model
Positive	89.452	90.247	91.176	93.214	97.032
Negative	84.326	86.675	88.675	89.956	94.876
Neutral	79.677	82.753	84.783	86.309	92.689

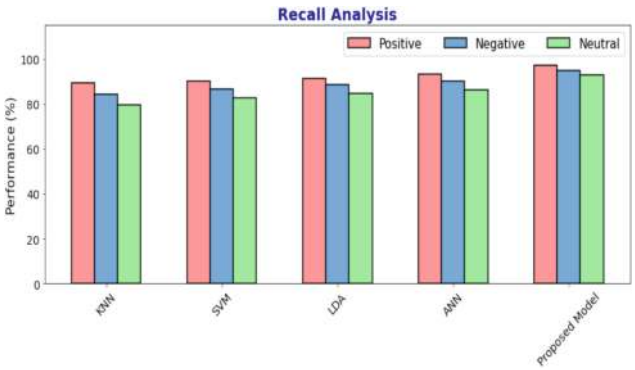


Fig. 5. Recall analysis for proposed model with other models.

The recall performance of the proposed model is compared with four traditional classifiers (KNN, SVM, LDA, and ANN) in three sentiment classes (Positive, Negative, and Neutral) in Table 3 and Fig. 5. The proposed model consistently outperforms all baseline methods, achieving the highest recall in every category. It attains a recall of 97.032% for the Positive class, surpassing ANN's 93.214% and significantly outperforming KNN's 89.452%. For the Negative class, it achieves a recall of 94.876%, compared to 89.956% for ANN and 84.326% for KNN. Similarly, in the Neutral class, the model records a recall of 92.689%, exceeding ANN's 86.309% and KNN's 79.677%. These findings show the increased capacity of the suggested method to true positively recognize the accurate cases of all the categories, which attests to the high levels of robustness and reliability of the method in real-life classification.

Table 4. F1-score analysis for proposed model with another model.

Classes	KNN	SVM	LDA	ANN	Proposed model
Positive	87.245	89.404	90.707	92.367	96.324
Negative	83.187	85.838	87.476	89.751	94.516
Neutral	78.946	81.343	84.765	85.563	92.019

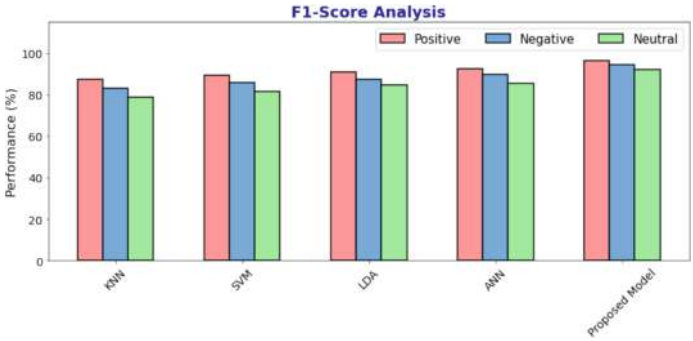


Fig. 6. F1-score analysis for proposed model with other models.

Table 4 and Fig. 6 compare the F1-scores achieved by various models (KNN, SVM, LDA, ANN, and the proposed model) across three sentiment classes: positive, negative, and neutral. The proposed model consistently outperforms all baseline approaches, indicating superior classification capability. For the positive class, the suggested model attains an F1-score of 96.324, which is significantly higher than those of KNN (87.245), SVM (89.404), LDA (90.707), and ANN (92.367), demonstrating a strong balance between precision and recall in accurately identifying positive instances. Similarly, for the negative class, the proposed method achieves an F1-score of 94.516, outperforming KNN (83.187), SVM (85.838), LDA (87.476), and ANN (89.751), indicating effective recognition of negative sentiments. In the neutral class, which is generally the most challenging due to overlapping features, the proposed model achieves an F1-score of 92.019, surpassing the next best model (ANN) by more than 6 percentage points. These results highlight the effectiveness and reliability of the proposed model in performing multiclass sentiment classification with high accuracy and consistency.

Table 5 and Fig. 7 introduce comparative research of the classification performance with the proposed and several available works on EEG-based mental workload estimation. This literature covers a variety of years, datasets, and neural structures. As an illustration, [23] (2020) used a BLSTM-LSTM classifier on the STEW dataset and obtained a rate of 82.57%. A hybrid Bi-LSTM and ResNet model was used in [33] (2023) on the same data, with an accuracy of 90.64%. In the same way, [34] (2021) and [35] (2020) applied Conv-LSTM and LSTM for processing the datasets of laboratory experiments, respectively, and reported the accuracy of 93.75% and 90.8%. Consequently, the suggested model, which was trained on the SEED dataset and was based on the

Table 5. Comparison with other models.

Paper	Classifier	Metric	Dataset	Year	Metric value (%)
[19]	BLSTM-LSTM	Accuracy	STEW	2020	82.57
[20]	Bi-LSTM and ResNet	Accuracy	STEW	2023	90.64
[21]	Conv-LSTM	Accuracy	Lab experiment	2021	93.75
[22]	LSTM	Accuracy	Lab experiment	2020	90.8
Our model	CSSL-TE	Accuracy	SEED	–	97.45

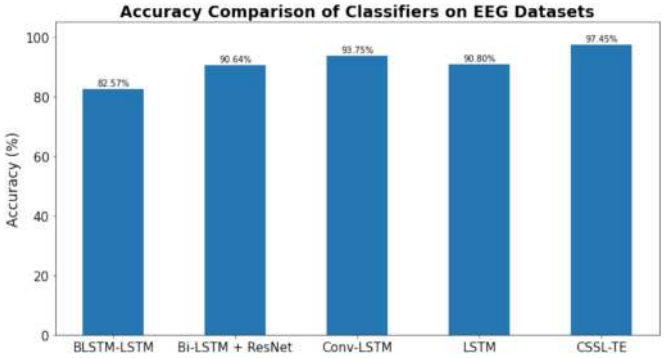


Fig. 7. Comparison with existing studies.

Contrastive Self-Supervised Learning and Transformer Encoders (CSSL-TE), achieves an impressive accuracy of 97.45% as compared to all other previous methods. This is better due to the fact that the Transformer can extract long-range temporal dependency and the self-supervised learning method is effective in extracting the invariant EEG representations.

5 Conclusion

The paper introduces a new deep learning model that uses a combination of CSSL and Transformer Encoder (TE) to predict mental workload reliably and without bias on raw EEG signals, especially in MS patients. The model learns to provide the complex brain patterns in the form of invariant representations using unlabeled data by contrastive pre-training, which does not require extensive manual annotation. The Transformer Encoder additionally assists in better modeling time-dependence as well as contextual information amongst EEG sequences. Experimentally, the proposed CSSL-TE model is more successful than the traditional classifiers like KNN, SVM, LDA and ANN in all the performance metrics including recall, precision, F1-score and accuracy, by measuring the model using the SEED dataset. It has the overall accuracy of 97.45. The findings indicate that attention-based models and self-supervised learning can be effective to achieve

scalable, interpretable and clinically viable EEG-based systems of workload classification. This approach shows a lot of promise for adaptive human-computer interaction and real-time cognitive monitoring in neuroergonomic and assistive health care settings. More studies can be done in future that are based on this study but includes cross-modal data, longitudinal, and domain adaptation to estimate personalized workloads among different clinical populations.

References

1. Hussain, I., Park, S.J.: Quantitative evaluation of task-induced neurological outcome after stroke. Brain Sci. **11**, 900 (2021). [Google Scholar] [CrossRef]. https://doi.org/10.3390/brainsci11070900
2. Hussain, I., Hossain, A., Jany, R., Bari, A., Uddin, M., Raihan, A., Kamal, M., Ku, Y., Kim, J.: Quantitative evaluation of EEG-biomarkers for prediction of sleep stages. Sensors **22**, 3079 (2022). [Google Scholar] [CrossRef]. https://doi.org/10.3390/s22083079
3. Hussain, I., Young, S., Kim, C.H., Benjamin, H.C.M., Park, S.J.: Quantifying physiological biomarkers of a microwave brain stimulation device. Sensors **21**, 1896 (2021). [Google Scholar] [CrossRef] [PubMed]. https://doi.org/10.3390/s21051896
4. Kakkos, I., Dimitrakopoulos, G.N., Sun, Y., Yuan, J., Matsopoulos, G.K., Bezerianos, A., Sun, Y.: EEG fingerprints of task-independent mental workload discrimination. IEEE J. Biomed. Health Inform. **5**, 3824–3833 (2021). https://doi.org/10.1109/JBHI.2021.3085131
5. Maimon, N.B., Molcho, L., Intrator, N., Lamy, D.: Single-channel EEG features during n-back task correlate with working memory load. arXiv 2020, arXiv:2008.04987. [Google Scholar]. https://doi.org/10.1109/LifeTech52111.2021.9391963
6. Shan, X., et al.: A revised Hilbert-Huang transformation to track non-stationary association of electroencephalography signals. IEEE Trans. Neural Syst. Rehabil. Eng. **29**, 841–851 (2021). https://doi.org/10.1109/TNSRE.2021.3076311
7. Kumar, A., Karmakar, S., Agarwal, I., Pal, T.: Mental workload classification with one-dimensional CNN using fNIRS signal. In: International Conference on Pattern Recognition and Machine Intelligence, pp. 746–755. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-45170-6_78
8. Nour, M., Öztürk, Ş, Polat, K.: A novel classification framework using multiple bandwidth method with optimized CNN for brain–computer interfaces with EEG-fNIRS signals. Neural Comput. Appl. **33**, 15815–15829 (2021). https://doi.org/10.1007/s00521-021-06202-4
9. Hasan, M., Mahmud, M., Poudel, S., Donthula, K., Poudel, K.: Mental workload classification from fnirs signals by leveraging machine learning. In: 2023 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE (2023). https://doi.org/10.1109/SPMB59478.2023.10372653
10. Chu, H., Cao, Y., Jiang, J., Yang, J., Huang, M., Li, Q., Jiao, X.: Optimized electroencephalogram and functional near-infrared spectroscopy-based mental workload detection method for practical applications. Biomed. Eng. Online **21**(1), 9 (2022). https://doi.org/10.1186/s12938-022-00980-1
11. Guan, K., Zhang, Z., Chai, X., Tian, Z., Liu, T., Niu, H.: EEG based dynamic functional connectivity analysis in mental workload tasks with different types of information. IEEE Trans. Neural Syst. Rehabil. Eng. **30**, 632–642 (2022). https://doi.org/10.1109/TNSRE.2022.3156546
12. Guan, K., Zhang, Z., Liu, T., Niu, H.: Cross-task mental workload recognition based on EEG tensor representation and transfer learning. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 2632–2639 (2023). https://doi.org/10.1109/TNSRE.2023.3277867

13. Pušica, M., Caiazzo, C., Djapan, M., Savković, M., Leva, M.C.: Towards practical deployment: subject-independent EEG-based mental workload classification on assembly lines. In: 2024 11th International Conference on Electrical. Electronic and Computing Engineering (IcETRAN), pp. 1–4. IEEE (2024). https://doi.org/10.1109/IcETRAN62308.2024.10645152
14. Kingphai, K., Moshfeghi, Y.: On channel selection for EEG-based mental workload classification. In: International Conference on Machine Learning. Optimization, and Data Science, pp. 403–417. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-53966-4_30
15. Wang, Z., Ouyang, Y., Zeng, H.: ARFN: An attention-based recurrent fuzzy network for EEG mental workload assessment. IEEE Trans. Instrum. Meas. (2024). https://doi.org/10.1109/TIM.2024.3369143
16. https://www.kaggle.com/datasets/phhasian0710/seed-iv
17. Şentürk, Ü.: Advanced EEG signal processing with deep autoencoders and hybrid Mamba classifier for accurate classification of chronic neuropathic pain etiologies. Ain Shams Eng. J. **16**(7), 03436 (2025). https://doi.org/10.1016/j.asej.2025.103436
18. Zhou, Y., et al.: Cross-task cognitive workload recognition based on EEG and domain adaptation. IEEE Trans. Neural Syst. Rehabil. Eng. **30**, 50–60 (2022). https://doi.org/10.1109/TNSRE.2022.3140456
19. Chakladar, D.D., Dey, S., Roy, P.P., Dogra, D.P.: EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm. Biomed. Signal Process. Control **60**, 101989 (2020). https://doi.org/10.1016/j.bspc.2020.101989
20. Shao, S., Han, G., Wang, T., Lin, C., Song, C., Yao, C.: EEG-based mental workload classification method based on hybrid deep learning model under IoT. IEEE J. Biomed. Health Inform. (2023) [Google Scholar] [CrossRef]. https://doi.org/10.1109/JBHI.2023.3281793
21. Gupta, A., Siddhad, G., Pandey, V., Roy, P.P., Kim, B.-G.: Subject-specific cognitive workload classification using EEG-based functional connectivity and deep learning. Sensors **21**, 6710 (2021). [Google Scholar] [CrossRef] [PubMed] https://doi.org/10.3390/s21206710
22. Kwak, Y., Kong, K., Song, W.-J., Min, B.-K., Kim, S.-E.: Multilevel feature fusion with 3D convolutional neural network for EEG-based workload estimation. IEEE Access **8**, 16009–16021 (2020). https://doi.org/10.1109/ACCESS.2020.2966834
23. Privalov, A.N., Bogatyreva, Y.I., Lapina, M.A., Lapin, V.G., Mysina, Y.A.: Decision support information system for patient treatment procedures in hospital. Ceur Workshop Proc. **2914**, 441–448 (2021)

Issues and Challenges in Multimodal Sentiment Analysis

Priya Singh[1], Rajesh Lavania[1], Mahfooz Alam[2(✉)], Mohammad Shahid[3], and Vitalii Lapin[4]

[1] Department of Computer Science, Institute of Engineering and Technology, Dr. B. R. Ambedkar University, Khandari, Agra, India
[2] Department of MCA, G. L. Bajaj Institute of Technology and Management, Greater Noida, India
mahfoozalam.amu@gmail.com
[3] Department of Commerce, Aligarh Muslim University, Aligarh, India
[4] North Caucasus Federal University, Stavropol, Russia

Abstract. Multimodal sentiment analysis (MSA) integrates multiple data sources such as text, audio, and visual modalities to achieve a more accurate understanding of sentiment. While this approach enhances sentiment prediction, it also introduces several challenges, including data fusion complexity, modality synchronization, data heterogeneity, and interpretability issues. This paper explores the primary issues and challenges in multimodal sentiment analysis, reviews recent advancements, and provides insights into potential solutions and future directions. However, to detect irony and other exaggeration, a typical finding from text data is. Multimodal analysis includes rich visuals and sound information in addition to text. It also uses fusion analysis to derive the implicit polarity (positive, neutral, and negative) of the feeling more accurately.

Keywords: Sentiment analysis · Deep learning · Artificial intelligence · Multimodal sentiment analysis · Natural language processing

1 Introduction

Sentiment analysis is an essential area in natural language processing (NLP) and artificial intelligence (AI), often used in applications like opinion mining, customer review analysis, and human-computer interaction. In the past, the focus of sentiment analysis was strictly on linguistics; yet, people communicate in a myriad of ways that include text, speech, gestures, and even facial expressions. The goal of multitasking sentiment analysis (MSA) is to improve the accuracy of sentiment classification by using multiple inputs [1–3]. The application of MSA is hindered by multiple practical and theoretical challenges, despite the advantages the approach offers. Over the past two decades, the increased strength of AI technologies has led to an increase in the effectiveness of multimodal analysis as compared to unimodal analysis. This, in turn, has changed the workings of sentiment analysis by incorporating audio, text, and video. The rapid growth

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 661–672, 2027.
https://doi.org/10.1007/978-3-032-27834-0_58

of social networks like Facebook and YouTube has led to a surge in the production of video content, capturing people's thoughts, thus creating more data for analysis. The levels of analysis may include document level, aspect level, and even sentence level [9].

The levels of sentiment can be explained further through document-level analysis. A document can be categorized as positive, negative, or neutral in its entirety. This classification depends on the intensity of the author's sentiments regarding a particular subject, such as a specific product or service. It is important to note that documents that assess or contrast multiple items are unsuitable for this type of classification. Furthermore, document–level classification is most effective when the document is authored by a single individual. The techniques of sentiment analysis are classified as deep learning, machine learning, and ensemble learning, and the process of these techniques is shown in Fig. 1.

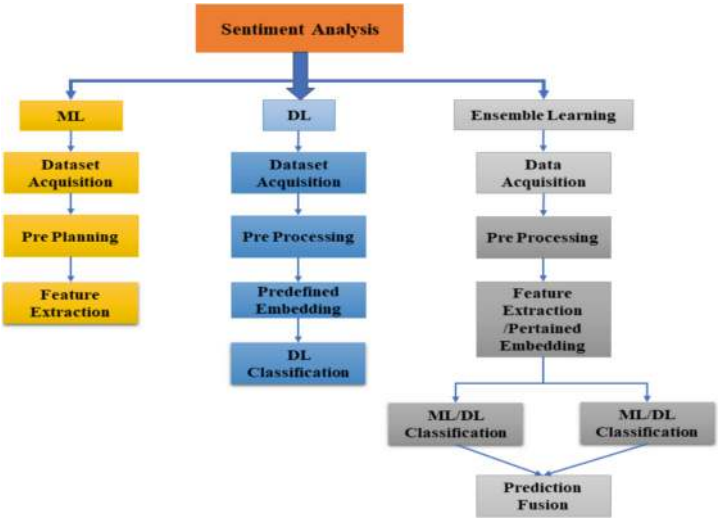


Fig. 1. Classification of Sentiment Analysis.

The remainder of the paper is structured as follows: Sect. 2 discusses the key challenges of multimodal sentiment analysis. Section 3 outlines the recent advancements in different technologies which are used in sentiment extraction. Section 4 sets forth the future scope of sentiment analysis, which encompasses the uses of sentiment analysis in upcoming social media networks. All these points are concluded in the paper.

2 Key Challenges in Multimodal Sentiment Analysis

In this section, a representation of key challenges of multimodal sentiment analysis is explained. The different multimodal sentiment analysis, such as data heterogeneity, data fusion complexity, synchronization of different modalities, identifying missing modalities in the real world, and the use of explainable AI.

2.1 Data Heterogeneity

Different modalities of data, such as text, audio, and visuals, require distinct methods of preprocessing and feature extraction. Text data needs syntactic and semantic analyses; audio data needs prosodic and spectral analyses; and visuals relate to facial recognition and expression detection. The primary issue is how to combine all of these diverse types of data. Supervised machine learning is applied for the proposed sentence-level and document-level sentiment analysis for heterogeneous datasets. The text documents of the different datasets are classified into two main types: subjective and objective.

2.2 Challenges of Data Fusion

Early or feature-level fusion, late or decision-level fusion, feature space manipulation-based fusion, and hybrid fusion are all ways to combine different modalities, which need to be done carefully. The methods of multimodal fusion are typically divided into early fusion and late fusion. Early fusion is achieved by merging features from different modalities into a single, unified input, while late fusion builds independent models for each modality and combines the outputs at a later stage.

Poorly designed fusion mechanisms can introduce redundancy, noise, or loss of critical information, impacting model performance. By combining the modality and features of each modality and merging them at the input part, which creates a joint representation of both, the resultant classification sentiment analysis will be performed. In EF-LSTM (early fusion) model, data is input by early fusion from different nodes of data [4]. Late fusion works to separate the modals of multi-language sentiments, the categorization results of late fusion depend on the identification accuracy of unimodal data [5].

Classifiers that have been trained for each modality, including text, audio, and visual, are applied to generate final sentiment predictions through a majority voting process. This strategy resembles an ensemble learning technique, as it amalgamates the strengths of the individual classifiers. As a result, it establishes a standard for performance assessment in comparison to more advanced and integrated multimodal strategies. The features obtained from each modality are merged at the final hidden state and subsequently processed through a classifier to determine the sentiment. This method facilitates the direct integration of features before classification, allowing the model to develop a unified representation of the modalities. The introduction of an attention layer facilitates the dynamic evaluation of the importance of features across various modalities. This capability allows the model to prioritize the most relevant features at each stage of the sentiment analysis process, thereby improving the integration of multimodal data [6].

2.3 Modality Synchronization

The different spatial and temporal resolutions of individual data streams create a unique challenge. Integrating text, audio, and video is a challenge even without considering data source sampling rates. Not having precise synchronization between streams can lead to inaccurate readings. The challenge of synchronization, termed as ‘modality synchronization’, is compounded in scenarios that involve audio and video because of incompatible

data alignment fusion. Different streams have different texts that accompany and represent them. The alignment between spoken text and written text is hard to achieve since both speech and text, as different modalities, are likely to have varying lengths.

2.4 Missing or Noisy Modalities

Poor audio, text, or even visual input in the form of occlusions can lead to missing unreliable modalities, which in turn can lead to degraded model performance. The challenge of maintaining model performance amid incomplete multimodal data is intensely complex. Removing a single modality during a model performance test introduces a shift in the input data distribution compared to the trained model, which risks the performance of the model mapping function. The unpredictable nature of the modalities in the real world emphasizes modality fidelity, thus limiting the applicability of earlier multimodal representational learning approaches. Various networks specific to each modality are employed to learn distinct representations. With respect to complete modalities text, audio, and vision "and the missing modalities, which include (text), (audio), (vision), (text, audio), (text, vision), and (audio, vision) the utilization of multimodal fusion networks that feature a uniform structure to capture representations for both complete and missing modalities. To ensure the effective transfer of semantic knowledge from complete modalities, all three constraints are used to align the representations of missing and complete modalities. These include geometric contrastive loss for constructive learning at the sample level, distribution distance loss to modify the representation distributions, and sentiment semantic loss to provide supervision based on sentiment labels [7].

2.5 Interpretability and Explainability

Deep learning models for MSA, such as transformers and recurrent neural networks (RNNs), often operate as black boxes. Understanding how different modalities contribute to sentiment predictions remains a challenge, limiting trust and interpretability in critical applications like healthcare and finance. Interpretability methods highlight the classification of explanations based on their generation timing, distinguishing between those created automatically during the prediction process (pre-hoc) and those that necessitate additional processing after a prediction is made (post-hoc). Directly interpretable or self-explanatory methods yield explanations in tandem with predictions, drawing on the model's output at the time of prediction. Global self-explaining models include decision trees and rule-based models, whereas local self-explaining models are characterized by attention-based approaches [8]. A Challenge to articulate which modality-text, audio, or video had a more pronounced effect on sentiment classification. For example, one might consider whether the video's facial expressions conveyed more information than the accompanying text or if the tone of voice was more impactful than the spoken words.

2.6 Scalability and Computational Cost

MSA models are computationally expensive, requiring high processing power for real-time applications. Training deep multimodal models demands substantial hardware

resources, making it difficult to deploy on edge devices or in low-resource environments. Multimodal systems operate with a variety of data streams, each possessing its own level of complexity and volume. Text data can encompass long–form written content, audio data may involve the processing of large segments of spoken language, and video data requires the handling of high–dimensional data, including frames or sequences of images. Therefore, the total data volume processed in multimodal systems can far surpass that of unimodal sentiment analysis.

The techniques of pruning, which is the elimination of unnecessary neurons or layers, and quantization, which is reducing the precision used for weights and activations, can alleviate the problem of scalability. These techniques, pruning and quantization, help to improve the level of accuracy, reduce the computational cost, and the smaller overall size of the models. This translates to improved inference speed. In real-time use cases such as virtual assistants or self-driving cars, performing sentiment analysis at the edge, for example, on smartphones or cameras, shifts the workload from centralized data centers and improves latency. This, however, requires the models to be telemetry agnostic.

2.7 Dataset Limitations and Biases

The rest of the existing multimodal datasets suffer from being smaller in size, unbalanced, or domain-specific, which, in turn, hinders generalization. Datasets can be biased, for example, culturally, where sentiment expression is biased towards a certain gender, which can lead to unfair or inaccurate sentiment predictions. A sentiment analysis system can adopt biases of the training data, and in multimodal sentiment analysis, such as bias from expression data, can skew the predictions. This results in inequitable sentiment classifications, particularly when evaluating demographic factors such as gender, race, or age within multimodal datasets.

3 Recent Advancements

Researchers have explored various solutions to address MSA challenges, which are: Self-Supervised Learning: A novel self-supervised learning framework utilizing auto encoders is introduced to enhance the model's capacity to extract sentiment information that is both modality-specific and modality–consistent, here by providing a thorough and distinct perspective on multimodal data. The model could utilize a self-supervised task like masked language modeling (e.g., BERT) or next-sentence prediction. Such tasks facilitate the model's ability to develop robust textual representations from large datasets without the necessity of sentiment labels. Methods like contrastive learning have improved multimodal representations.

Within the context of speech, self-supervised learning (SSL) can entail the acquisition of representations by forecasting missing audio segments or synchronizing speech with a written text. The various approaches, like contrastive predictive coding(CPC), are applicable for the given purpose. Multi-task learning is designed to improve the generalization performance of multiple interrelated tasks by identifying the knowledge available from a variety of tasks.

3.1 Graph Neural Networks (GNNs)

The GNN has shown promise in generating results for text classification, multi-label recognition, and various multimodal tasks. For text classification, a new neural network called GNN and its variants have rapidly advanced, outperforming traditional methods such as TextGCN (Graph Convolutional Network), Tensor GCN, and TextLevelGNN.

GNNs help model complex relationships between modalities. The representation of a single image–text post failed to encompass the broader co-occurrence dynamics of the datasets. In GNN, the use of Multi-channel GNNs with Sentiment–awareness (MGNNS) to enhance sentiment detection in image text contexts [11, 12]. The working of detecting sentiment analysis in multimodal formats, especially works with image–text pairs found in social media posts, is fraught with challenges. The issues of image–text misalignment and the imperfections present in social media content, such as causal language, spelling mistakes, and a lack of proper punctuation, create significant obstacles for the accurate representation of multimodal data for sentiment detection. To overcome these obstacles, researchers have developed various networks for multimodal sentiment analysis, including the Hierarchical Semantic Attentional Network (HSAN) and the Multimodal Deep Semantic Network (MDSN). Furthermore, the introduction of models like the Co-Memory network (Co-Mem) and the Multi-View Attentional Network (MVAN) has incorporated memory networks to enhance the interaction between different modalities [13, 14].

3.2 Multi-Task Learning

Across different modalities on a large amount of data, which is not labeled, Self-supervised training can be used for pre-training the model. There are different ways which show that the idea is to anchor SSL to learn a general representation that can be adapted for multiple tasks. These are Image modals (In this task, the ability to see, SSL techniques like Contrastive learning, which help the model to learn strong image presentation, and for tasks like object detection, sectionalisation, and categorization, this strong presentation can be used.) and Text modals (Using Self-supervised learning, detailed presentation of language is learned by the model, and for various downstream tasks like machine translation, emotion analysis, this rich presentation can be established). Shared representations across different sentiment tasks enhance learning efficiency.

Multi-tasking learning aims to exploit the useful information found in multiple related tasks to increase the generalization performance of the model. On the basis of single-task and multi-task models, learning gives the three key advantages: one is that the model's performance can be enhanced through the sharing of information, another is to exhibit improved generalization capabilities, and the last is necessitates only one comprehensive model, thereby reducing overall model complexity [16]. Due to these advantages, a widespread implementation of multitasking learning in MSA is possible. Multi-task networks apply parameter sharing techniques to optimize the model's results; the parameters are categorized into hard parameters and soft parameters sharing. Hard parameter sharing involves sharing of multiple tasks under certain hidden layers of the network, making predictions for various tasks before coming to the output stage. Whereas soft parameter sharing utilizes independent models for different tasks [17].

As the earlier studies show, sentiment analysis involves heterogeneous data, that data contains low-level features which can identify a specific insight into individual modalities, as well as high-dimensional features which encapsulate global semantic information across widespread modalities. The model creates challenges in balancing the extraction of intra-modal features with the fusion of inter-modal information. To mitigate this issue, the application of a multi-task learning methodology that prioritizes the dominant modality can enhance the stability of the sentiment analysis results. Cross-modal attention works as to adopt cross-modal multi-head attention to enhance information interaction between modalities [18].

3.3 Explainable AI (XAI) Methods

In SA (sentiment analysis), explainability concerns the ability to clarify the decision–making mechanisms of sentiment models, thus making their predictions more transparent and understandable. This shows a significant challenge, particularly with complex models like DL techniques, which are considered "black boxes". This issue of clarity makes it difficult to comprehend the reasoning behind the assignment of specific sentiments. A pitfall in explainability can result in misbelieve and a lack of trust in the results of sentiment analysis.

Attention mechanisms and saliency maps are being used to interpret MSA models. Explainable AI (XAI) models are necessary for advancements in interpretability and transparency of AI systems, especially in intricate various tasks MSA. It involves the assessment of sentiment among various modalities, such as text, images, audio, or videos, which are frequently utilized in deep learning models, and can be a challenging task to interpret the data. The adoption of XAI techniques enables users to summarize the reasoning behind various model decisions, which produces trust and ensures fairness and accountability. In the field of multimodal sentiment analysis, a LIME can produce explanations for a specific prediction by the inputs and analyze how the model's outputs are changed. For instance, a multimodal sentiment analysis can process both text and image, a LIME can represent how the changing parts of the text or parts of the image affect the predicted sentiment.

Explainable AI is an emerging area which focuses on including transparency and interpretability to typical DNNs or transformer-based LLMs. XAI models are often referred to as black –box models due to the lack of ability to provide insight into their decision-making processes, or they offer explanations for their predictions and biases. XAI is used to collect a collection of processes and methods that empower human users to understand and trust the outputs created by ML algorithms. XAI shows its value in medical imaging research, human computer interaction (HCI), stock price prediction, and NLP tasks, which provide insights into ML models that work over application processes and human-centered approaches [19]. A comparative analysis of the different recent techniques is summarized in the following Table 1.

Table 1. A comparison of deep learning techniques.

Techniques	Self-supervised learning (SSL)	Graph neural networks (GNN)	Explainable AI (XAI)	Multitasking (MTL)
Objective	It learns representations of data from large, unlabeled multimodal big datasets	It stores complex data dependencies and relationships in multimodal datasets	It makes sentiment analysis models interpretable and understandable	Training of a single model on multiple relationship tasks (e.g., text, speech, images)
Advantage	There is no requirement for labeled datasets, can scale up with large amounts of unlabeled multimodal data	It is an ideal way for relational data, which captures dependencies across different modalities	It generates belief and transparency in multimodal sentiment predictions	It learns shared data representations b/w different modalities for better performance
Primary uses	It is specially used in pre-training models on multimodal data (e.g., text and image) without labels or unlabeled data	Its uses are in Social media analysis, recommendation systems, multimodal and cross-lingual sentiment relationships	It is used in legal, healthcare, or financial applications requiring a model where trust and transparency are required	It is used for the classification of multimodal sentiment from text, speech, images, audio, and video
Challenges	To design pretext tasks for multimodal sentiment analysis, the a risk of trivial learning of the datasets	It is typical to handle scalability for complex multimodal graphs, over-smoothing	In these difficult to explain complex multimodal models, for balancing the explainability with performance	In this, it is difficult to balance multiple objectives for conflicting tasks or gradients
Performance	It can learn strong representations from different data sources without needing annotations	It gives a better relational structure of sentiment across multiple modalities	It provides trust, enabling real-world deployment of AI systems in sensitive domains of sentiment analysis	It provides better generalization between the tasks to improve sentiment analysis between diverse input types

4 Future Directions

Future research should focus on or its works may prioritize the enhancement of both the accuracy and the scope of sentiment analysis techniques. It can include the development of sophisticated models that effectively store the emotional intensity of textual content, and which quantify sentiment across a variety of data, addressing the challenges posed by ambiguous and sarcastic texts, expanding the framework of cross-lingual sentiment analysis [20]. An end –to –end multimodal network for learning of data can generate and produce a relation between multimodal data learning and unimodal data learning [10]. The various future scopes of the SA in multimodal techniques are classified in Fig. 2.

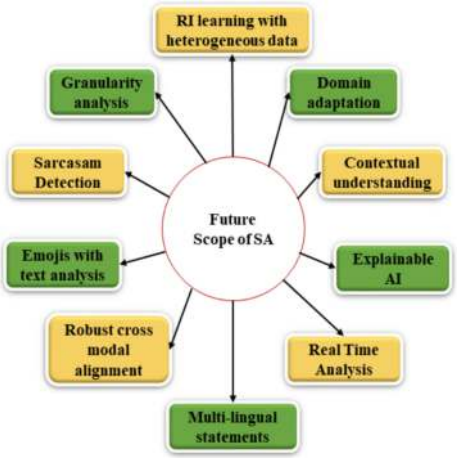


Fig. 2. Future scope of sentiment analysis.

Fine-grained sentiment analysis is a new advancement in sentiment analysis models, which primarily classifies sentiment into three broad categories: positive, negative, and neutral. There remains a necessity to improve fine-grained sentiment analysis, encompassing various emotional influences, including very positive, positive, neutral, negative, and strongly negative. Researchers can refine diverse deep learning architectures and methodologies to execute fine-grained sentiment analysis. The diverse methodologies employ a hierarchical attention network capable of capturing sentiment articulated in distinct sections of a text at many levels of granularity [20].

Sarcasm is defined as the employment of remarks that convey a message or attitude that is entirely contrary. It is designed to identify or irritate an individual for diverse comedic purposes, utilising deep learning technologies. The presence of sarcasm in a document diminishes the accuracy of sentiment analysis relative to other straightforward methods, which are unable to identify sarcasm across diverse sentiments [23].

Emojis help social media users to share their feelings easily with less text content. Emoji's were found to help the improvement of overall sentiment analysis scores. While emoji's characters are used for expressing negative and positive opinions, which helps

enhance sentiment expressions through large opinion of sentiment scores of positive feelings more than negative feelings in the analysis of data [24].

Multimodal sentiment analysis is capable of developing strict cross-modal alignment techniques that are based on traditional analysis of text. Wherever on the an enhancement of multimedia, there is a need for sentiment analysis to be extended in the midst of multimedia, such as videos, images, and audio. This technique will permit the extraction of more accurate data by taking into account the interactions among different modalities [21].

Cross-Lingual sentiment analysis is the use of sentiment models that are trained on english text. There is a requirement for such models increasing which can linguistically analyze sentiment. the implementation of cross-lingual sentiment to enhance the understandability of sentiments which are represented in various languages, thereby expanding the accessibility of SA to a wider audience. An effective approach involves pre-training these models on big multilingual datasets and, as a result, fine-tuning them for a particular sentiment analysis task in different languages [20].

Real-Time sentiment analysis helps to do in the creation of larger, unbiased, and diverse multimodal datasets. RTSA becomes a critical component with the enhancement of social media data with instant messaging platforms. different companies can use SA to handle customer feedback, and they quickly respond to customers [21].

The generation of efficient and interpretable MSA models for real-time applications is necessary. The emotional and text expressions with emoji are related to sentiment and can change remarkably from those in different domains. Therefore, there is a need for sentiment analysis that is specific to every domain, and permits the adaptation of language and sentiment to the unique properties of that area [22].

Enhancing computational efficiency to enable deployment in resource-constrained environments. Explainable sentiment analysis aims to provide an explanation for the sentiment analysis result, improving the transparency and interpretability of the sentiment analysis model.

Domain adaptation sentiment analysis, domain adaptation is typical due to considerable differences in text data, which occur across various domains, languages, and writing styles. models designed for sentiment analysis often struggle with various domains in different fields, changes in vocabulary, different writing techniques, and changes in sentiment expressions in the climate of social media, product feedback, and news articles, which can negatively impact their accuracy of their results [22].

Contextual understanding is an intrinsic part of SA, Demanding the recognition of the subtle nuances that shape the sentiments shown through a text. This presents a notable challenge, as sentiments are primarily influenced by context and can be affected by elements such as sarcasm, negation, irony, and ambiguous language [21].

The efficiency of sentiment analysis models to correctly interpret confusions and sarcastic texts is notable. researchers may implement reinforcement learning techniques to train the models and allow them to effectively address the challenges posed by such nuanced language. The process contributes to the advancement of models that remember data from the feedback system and enhance their predictive capabilities [20].

5 Conclusion

Sentiment Analysis involves the classification and identification of emotional emojis or attitudes expressed in a text, which can be a positive, negative, or neutral tone of data. It is essential to comprehend the public emotions, customer feedback, and analysis of market trends. It enables organizations and businesses to make for formal land information and enhances user experiences. Multimodal sentiment analysis has the potential to revolutionize human-computer interaction and opinion mining. However, significant challenges remain in data fusion, synchronization, interpretability, and scalability. Addressing these challenges will pave the way for more robust and efficient sentiment analysis systems in the future. And lastly, Explainable AI sentiment analysis will be more important as the demand for transparency and interpretability for ML models continues to grow. The gathering of explainability will permit users to trust and understand the sentiment analysis results much better than other models' data.

References

1. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inform. Fusion **37**, 98–125 (2017)
2. Zadeh, A., Lim, Y.C., Morency, L.P., et al.: Multi-attention recurrent network for human communication comprehension. In: AAAI Conference on Artificial Intelligence (2018)
3. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2019)
4. Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., Morency, L.: Multiattention recurrent network for human communication comprehension. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pp. 5642–5649. AAAI Press (2018)
5. Zhu, L., Zhu, Z., Zhang, C., Xu, Y., Kong, X.: Multimodal sentiment analysis based on fusion methods: a survey. (2023)
6. Chen, Q., Dong, S., Wang, P.: Advanced multimodal sentiment analysis with enhanced contextual fusion and robustness (AMSA-ECFR): symmetry in feature integration and data alignment. Symmetry **16**(7), 934 (2024). https://doi.org/10.3390/sym16070934
7. Lin, R., Hu, H.: MissModal: increasing robustness to missing modality in multimodal sentiment analysis. Trans. Assoc. Comput. Linguist. **11**, 1686–1702
8. Arya, V., et al.: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. CoRR (2019). https://doi.org/10.48550/ARXIV.1909.03012
9. Liu, B.: Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies. Springer, Cham (2012)
10. Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis
11. Yang, X., Feng, S., Zhang, Y., Wang, D.: Multimodal sentiment detection based on multi-channel graph neural networks
12. Xu, N.: Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, pp. 152–154 (2017). https://doi.org/10.1109/ISI.2017.8004895
13. Yue, L., Chen, W., Li, X., Zuo, W., Yin, M.: A survey of sentiment analysis in social media. Knowl. Inform. Syst. 1–47 (2018)

14. Yang, X., Feng, S., Wang, D. and Zhang, Y.: Image-text multimodal emotion classification via multi-view attentional network (2020)
15. Zhang, Y., Li, J., Zhu, F., Cai, Y., Li, X., Rao, L.: Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning
16. Foggia, P., Greco, A., Saggese, A., Vento, M.: Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition. Eng. Appl. Artif. Intell. **118**, 105651 (2023)
17. Vandenhende, S., et al.: Multi-task learning for dense prediction tasks: a survey. IEEE Trans. Pattern Anal. Intell. Mach. **44**, 3614–3633 (2021)
18. Cai, Y., Li, X., Zhang, Y., et al.: Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Sci. Rep. **15**, 2126 (2025). https://doi.org/10.1038/s41598-025-85859-6
19. Mabokela, K.R., Primus, M., Celik, T.: Explainable pre-trained language models for sentiment analysis in low-resourced languages. Big Data Cogn. Comp. **8**(11), 160 (2024)
20. Tan, K.L., Lee, C.P., Lim, K.M.: A survey of sentiment analysis: approaches, datasets, and future research. Appl. Sci. **13**, 4550 (2023). https://doi.org/10.3390/app13074550
21. Pathak, U., Piyush, E.R.: Sentiment analysis: methods, applications, and future directions. Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) 11(2) (2023)
22. Jim, J.R., Talukder, M.A.R., Malakar, P., Kabir, M.M., Nur, K., Mridha, M.F.: Recent advancements and challenges of NLP-based sentiment analysis: a state-of-the-art review. Elsevier B.V
23. Tan, Y.Y., Chow, C.O., Kanesan, J., Chuah, J.H., Lim, Y.: Sentiment analysis and sarcasm detection using deep multi-task learning wireless. Pers. Commun. **129**, 2213–2237. https://doi.org/10.1007/s11277-023-10235-4
24. Elsalam, M.M., Gadallah, A.M., Hefny, H.A.: Sentiment analysis of text incorporating emojis: a machine learning approach. Int. J. Comp. Appl. (0975–8887) **184**(20) (2022)

Automated Diagnosis of Mental Disorders Based on the Analysis of Speech Biomarkers Using Machine Learning

Danil Trofimov[1(✉)], Maria Zapriy[2], Igor Kotilevets[1], and Daria Smirnova[3]

[1] Department of Digital Data Processing Technologies, MIREA—Russian Technological University, Moscow, Russia
danila.trofimov.0807@gmail.com

[2] Institute of Socio-Humanitarian and Digital Development of Medicine, Samara State Medical University, Samara, Russia

[3] Institute of Mental Health, University «Reaviz», Saint Petersburg, Russia

Abstract. The prevalence of mental disorders continues to grow worldwide, while diagnostic practice still relies heavily on subjective clinical assessment and time-consuming psychometric instruments. This study explores automated detection of mental disorders using speech by combining interpretable linguistic and acoustic biomarkers with deep learning methods. We employed a multiclass dataset covering five diagnostic categories of ICD-10 and developed two classification systems: an interpretable model based on handcrafted linguistic and acoustic features, and a convolutional neural network trained on Mel-frequency cepstral coefficient (MFCC) spectrograms. Both systems were evaluated using cross-validation and clinically relevant metrics. Experimental results indicate that speech-based models can discriminate between diagnostic groups with promising performance, suggesting their potential as rapid, objective aids for screening and monitoring in clinical settings. The proposed approach contributes a reproducible feature set and an end-to-end acoustic pipeline that together advance automated tools for mental healthcare.

Keywords: Mental disorders · Speech analysis · Voice biomarkers · Artificial Intelligence · Machine learning · Anxiety disorders · Depressive disorders · Emotional burnout

1 Relevance of the Research

1.1 Current Status of the Problem and Application of Speech Analysis

The global burden of mental disorders is increasing, and reliable, timely diagnosis remains a major challenge for health systems. Conventional clinical assessment and psychometric instruments remain central to diagnosis but often depend on subjective interpretation and specialist time, which limits scalability and early detection. Speech contains a rich set of measurable signals—prosodic, phonatory, articulatory and linguistic—that reflect cognitive and affective states. Empirical studies associate features such

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 673–684, 2027.
https://doi.org/10.1007/978-3-032-27834-0_59

as reduced speech rate and monotony with depressive disorders, altered fluency with anxiety, and changes in voice quality with certain psychotic and neurodegenerative conditions. According to the World Health Organization, approximately one in eight people worldwide currently suffer from mental disorders, which in absolute terms amounts to more than one billion people. Anxiety and depressive disorders are the most prevalent, with more than 300 million and 280 million cases, respectively. Among affective states: euthymia—38.8%, normal sadness—27.3%, emotional burnout—25.2%, depression—8.7% [1]. Despite these findings, research is hindered by heterogeneous datasets, inconsistent feature extraction protocols and limited attention to model interpretability, all of which reduce reproducibility and clinical applicability. Many high-performing models remain black boxes for clinicians, while simpler, interpretable pipelines often lack comparative evaluation across multiple diagnostic categories.

It has been established that acoustic and linguistic parameters of the voice can serve as biomarkers, allowing deviations related to mental health to be recorded. For example, slow speech and monotonous intonation may be biomarkers of a depressive states, while rapid speech and variability in timbre can be indicators of anxiety [2]. The development of artificial intelligence technologies and machine learning methods opens up new prospects for the automation of speech and voice analysis, creating conditions for the identification of hidden patterns associated with mental disorders.

The relevance of the study is determined by the steady trend towards an increase in the number of mental disorders on a global scale. According to the World Health Organization, the annual increase in cases of mental disorders ranges from 26 to 28% [3]. Traditional diagnostic approaches are mostly based on subjective interpretation of symptoms and require significant time and resources. The introduction of automation methods into this process, based on speech and voice analysis using machine learning technologies, opens up the possibility of accelerated detection of disorders and timely initiation.

To address these gaps, this work develops two complementary classification approaches: an interpretable pipeline built from handcrafted linguistic and acoustic features, and a convolutional neural network applied to MFCC spectrograms. We assess both systems on a multiclass corpus of speech recordings spanning five diagnostic category and report cross-validated performance using clinically relevant metrics. The goal is to evaluate the diagnostic separability provided by speech biomarkers and to deliver a reproducible, technically transparent workflow that may support screening and longitudinal monitoring in clinical practice. The scientific contribution of this work lies in the fact that the creation and implementation of a model capable of automating the analysis of speech characteristics will improve the effectiveness of the diagnosis and treatment of mental disorders. This, in turn, contributes to improving the quality of life of patients and reducing the socio-economic burden caused by the growing prevalence of these diseases.

1.2 Literature Review and Clinical Rationale

The primary stage of clinical work is differential diagnosis, aimed at distinguishing between mental disorders and normal, situational emotional reactions.

The principal approach remains the clinical method, which encompass collecting medical history, clinical interviews and monitoring of the patient's mental status.

Psychometric techniques employ standardized tests, questionnaires, and rating scales to provide quantitative evaluations of psychological traits and symptoms.

Laboratory approaches rely on the examination of biological samples and physiological markers, allowing somatic causes of mental symptoms to be ruled out and the correctness of the diagnosis to be confirmed [4, 5]. These include genetic analyses, hormonal and metabolic assays, immunological profiling, and toxicological testing.

Instrumental diagnostics provide visualization and assessment of brain function, allowing organic pathologies to be ruled out and functional disorders to be identified. These include neuroimaging, electroencephalography with evoked potentials, as well as computerized neuropsychological tests, polygraphy, Dopplerography, and analysis of vegetative indicators [6].

Contemporary machine learning technologies enhance the potential of established diagnostic tools, such as clinical interviews, laboratory data, and psychometric tests [1], and open up new diagnostic possibilities by enabling the analysis of linguistic and paralinguistic biomarkers, the user's digital phenotype, and data from wearable devices. These approaches accelerate the detection of disorders and help predict the course of diseases, but they require further validation and currently serve as an auxiliary tool alongside clinical assessment [7–10].

Machine learning technologies enable the processing of large datasets and the detection of subtle features that are difficult to capture using traditional methods, such as micro-fluctuations in speech tone or hidden behavioral patterns on social media. However, the implementation of such technologies comes with a number of limitations. These include the risk of sample bias arising when training models on English-language data, which reduces their universality of application in other linguistic environments, as well as problems with interpreting results and insufficient model validity, which causes caution on the part of the clinical community. These methods are rapid, objective, and relatively inexpensive but currently cannot substitute for clinical expertise, but their potential appears to be significant, especially in the Russian segment, where this area remains underdeveloped.

A comparative analysis of existing methods for diagnosing mental disorders is carried out taking into account parameters such as accuracy, sensitivity, specificity, validity, reliability, easy of use, speed of results, cost, and accessibility. Each parameter was assessed on a five-point scale, where the highest score corresponds to a pronounced advantage of the method, and the lowest score corresponds to its limitation. At the same time, Weighting coefficients ranging from 1 to 3 were applied into the analysis to take into account the relative substantial of the indicators: the minimum multiplier is taken as one, and the maximum multiplier is taken as three. The summary data of the comparative analysis are presented in Figs. 1 and 2.

In the comparative analysis, the upper estimate is set at 100 points. Psychometric techniques achieved 87.5 points, which is due to their standardization and high sensitivity to symptoms. However, accuracy, validity, and repeatability are reduced, and specificity suffers more due to the subjective self-assessment of respondents and the influence of external factors. Self-questionnaires rarely allow for an unambiguous classification

of mental state, since high scores on the depression scale can be recorded in anxiety disorders, requiring a clinical context for correct interpretation.

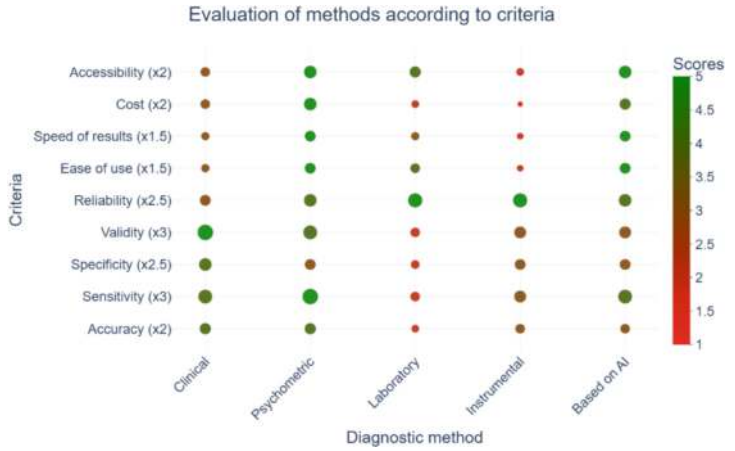


Fig. 1. Distribution of points according to criteria and methods.

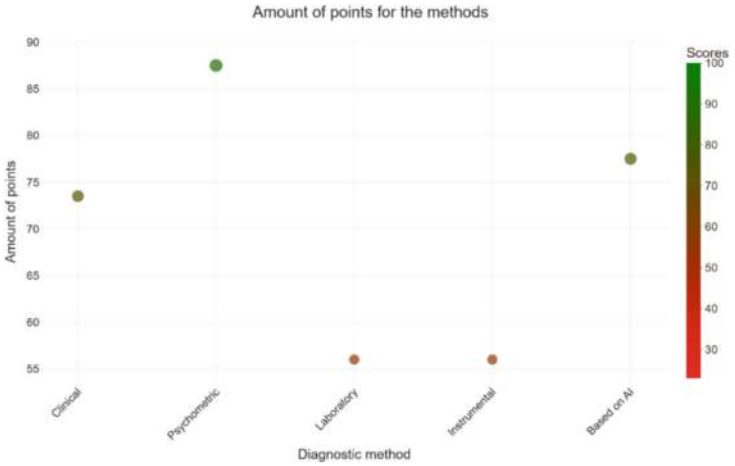


Fig. 2. The total number of points for each method.

Machine learning-based methods scored 77.5 points. They show promising sensitivity and allow for the detection of disorders at early stages, but the limited volume and quality of data reduce specificity, validity, and accuracy, in particular due to the lack of multi-class classifiers and insufficient elaboration of subtypes of disorders. These methods are fast, objective, and relatively inexpensive, but they are not yet capable of replacing clinicians and require further validation.

Impairments of cognitive and affective functions in mental disorders are reflected in the acoustic and linguistic characteristics of speech. Depression is characterized by monotonous, quiet and slow speech with pauses, simplified syntax and the use of vocabulary reflecting pessimism, hopelessness and anhedonia; anxiety is manifested by an increase in voice frequency, an acceleration of the pace of speech, tremor, speech about risks and threats; bipolar disorder in the manic phase is expressed by rapid, loud, expressive speech with puns and a leap of ideas, and in the depressive phase it is similar to unipolar depression; post-traumatic stress disorder can be accompanied by impoverished intonation, long pauses, and dysfluence when trying to describe a traumatic event and topics related to threat, fear, and guilt; schizophrenia is characterized by disorganized speech, neologisms, coherence breaks, and unusual associations with formal grammar preservation.

The choice of the final algorithm depends directly on the format of the input data, which by its nature reflects the biological, behavioral, cognitive, and social aspects of mental disorders and emotional dysfunctions. Data collection has its limitations due to the specificity of the field. The easiest to analyze and collect are the results of self-administered questionnaires and tests. The specificity also depends on the method of patient examination. While questionnaires are applicable in clinical and psychometric diagnostic methods, functional magnetic resonance imaginganalysis is relevant for instrumental methods, which affects the complexity and cost of obtaining such data.

Based on these premises, the following types of input data were identified:

1. Text data—this can include transcripts of clinical interviews, social media posts, personal diaries, dialogues, and more.
2. Neuroimaging—data on structural and functional changes in the brain based on magnetic resonance imaging or functional magnetic resonance imaging, electrical activity based on electroencephalography, and heart rate variability based on electroencephalography.
3. Physiological signals—eye movement data, voice biomarkers, facial expressions, other nonverbal signals.
4. Laboratory indicators—blood tests (cortisol, adrenocorticotropic hormone levels).
5. Demographic data and questionnaires—medical history, age, gender, social status [11], PHQ-9, DSM-5, GAD-7, DASS-21, Beck Depression Scale, Maslach Burnout Inventory questionnaire [1, 12].
6. Genetic and epigenetic data—analysis of SNPs (single nucleotide polymorphisms), gene expression [13], methylation [14].
7. Multimodal data—a combination of several input data formats. For example, analysis of video recordings of interviews (physiological signals and text data), family genetic analysis (demographic and genetic data).

To ensure analytical validity, the final dataset must meet several quality requirements: representativeness of the general population, class balance, normalization and scaling, minimal multicollinearity, feature relevance, absence of noise and outliers, completeness, homogeneity, and dimensional compatibility with the selected processing algorithm.

2 Description and Data Collection

Collecting audio data from respondents with mental disorders is fraught with a number of difficulties related to the quality of clinical diagnosis and the patient's readiness for the interview, as well as the technical characteristics of the recording equipment. The final dataset was compiled by two research groups using different protocols.

The first protocol «Mimic-speech monitoring of depression» was conducted as an open, non-interventional observational study. Three main groups were identified within the study: patients with mental disorders, a control group, and a fatigue subgroup formed on the basis of fatigue scales. The first audio recording was made at rest for one minute, and the second was made while presenting stimulus material according to the Karolinska Directed Emotional Faces design [15].

The second protocol «Spontaneous Speech Corpus» was conducted in an open, non-interventional observational format using audio-discursive tasks. Participants included patients from a psychiatric clinic and a control group from the general population. The study included three types of discursive tasks: picture narratives, personal reminiscences, and image-based instructions.

A comparative analysis of these studies reveals significant differences in design. First, the clinical domains differ: the first protocol focuses on the affective continuum, while the second focuses on the psychotic spectrum and thought disorders, with anxiety disorders almost entirely absent. Second, the sources of diagnosis are different: in the first protocol, the diagnosis is made by a psychiatrist upon inclusion, in the second, it is taken from the patient's medical history, and the control group has no diagnosis. Third, the assessment scales have limited overlap–only the Hamilton Rating Scale for Depression (HDRS) coincides, the other instruments are not directly compatible. Fourth, the data formats differ: emotionally provoked recordings versus audio discourse. Finally, the stratification of severity differs: the first protocol uses continuous HDRS scores, while the second uses a discrete gradation from 0 to 3.

To further assess data consistency, it is necessary to define the groups of conditions under study. Either the degree of confidence in the diagnosis or the interpretation of depression on the Hamilton scale for each sample is used as the second expert.

Due to differences in protocols, statistical measures of metric consistency were obtained (Table 1).

Based on the calculated indicators, the following conclusions can be drawn. The study of the first group demonstrates «substantial» agreement according to the Cohen's kappa coefficient, while the second group is characterized by «moderate» agreement. When combining the datasets, the Cohen's kappa coefficient tends to zero due to the absence of paired matches. High Cramer's V values within each individual study confirm the internal consistency of the experts, but when the data are combined, this consistency is weak, which is due to differences in the distribution of diagnoses. The KL divergence value is 1.709 nats, which corresponds to approximately 2.46 bits, indicating an informative discrepancy between the datasets.

Due to differences in study design and data collection methods, as well as statistical inconsistencies in the final samples, the overall accuracy of the classification may be reduced. When selecting a classification algorithm, class imbalance must also be taken into account.

Table 1. Statistical measures of consistency.

Metric	Research: «Mimic-speech monitoring of depression»	Research: «Corpus of spontaneous speech for psychiatry»	Between researches
Cohen's kappa	0,747	0,600	0,011
Chi-square	600,06, df = 182, $p \approx 4{,}2 \times 10^{-46}$	2337,54, df = 80, $p < 10^{-300}$	151,74, df = 4, $p \approx 8{,}6 \times 10^{-32}$
Kramer V	0,916	0,938	0,626
Kullback–Leibler Divergence	–	–	1,709 nats

Speech features were extracted in three ways: linguistically, acoustically, and based on mel-frequency cepstral coefficients (MFCC). The linguistic module processed the transcribed text, corrected errors, normalized the data, and extracted syntactic, morphological, semantic, and lexical-stylistic characteristics, including sentence complexity, pronoun usage, temporal consistency, modality, emotional coloring, and lexical diversity etc. The acoustic module recorded prosodic, vocal, formant, spectral, articulatory, and dynamic speech parameters, such as fundamental frequency, pauses, spectral characteristics, articulation clarity, and rhythm stability. The MFCC module automatically extracted spectral coefficients and their derivatives, which made it possible to analyze the dynamics of speech signals.

Two approaches were used for classification: the CatBoost algorithm, which works with interpretable linguistic and acoustic features, and a convolutional neural network for MFCC analysis. The quality of the models was evaluated using the weighted F1 measure, taking into account class imbalance. The most significant factors for recognition were the grammatical and semantic characteristics of the text, emotional markers, and voice parameters. Total of 227 and 96 features for different classifiers.

As a result, an optimal model was trained with 170 iterations, a maximum tree depth of 4, a gradient descent step of 0.1, and an L2 regularization coefficient of 5 for leaf weights. Figure 3 shows the change in the multi-class loss function during training.

The F1 measure is used as the main metric for algorithm quality, since the classes in the sample are unbalanced and the accuracy indicator may give overestimated values, reflecting only the success of recognizing dominant classes and ignoring minor ones. The F1 measure combines information about accuracy and completeness, is sensitive to class imbalance, and allows us to obtain both a separate score for each class and an average overall score, which is especially important when it is necessary to identify all categories evenly, not just the predominant ones.

Overall accuracy, class accuracy, reliability, weighted F1-measure, and weighted F1 are calculated. To evaluate the quality of the classification, a confusion matrix will be constructed and the 20 most important features that had the greatest impact on the model results will be displayed. The visualizations are shown in Fig. 4 and 5.

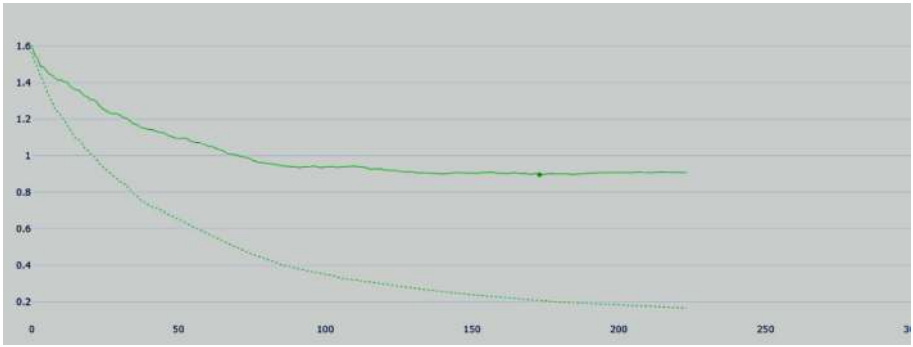


Fig. 3. Loss function during CatBoost classifier training, where the X-axis represents the number of iterations and the Y-axis represents the loss function value. The dotted line indicates the value for the training sample, and the solid line indicates the value for the validation method.

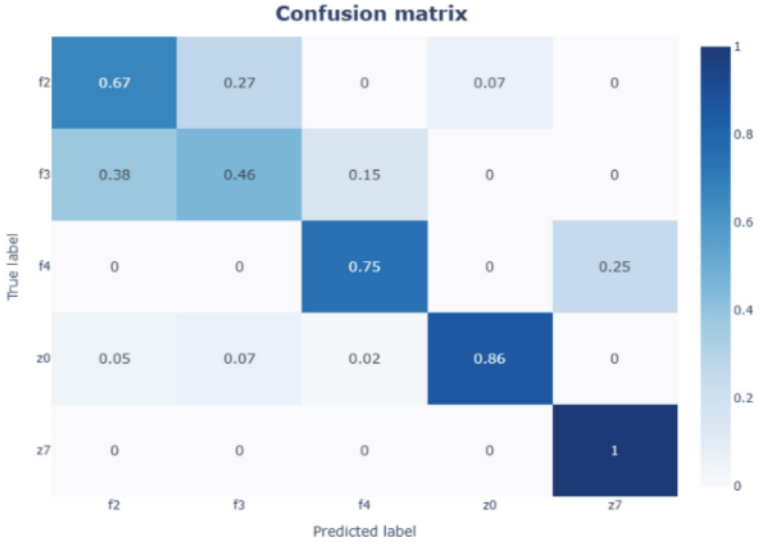


Fig. 4. Confusion matrix for evaluating the recognition quality of each class (CatBoost).

The most significant features for classification include the proportion of verbs and verb forms in the present and past tenses, articulation accuracy, the proportion of aggressive emotions, spectral uniformity, the average value of the eighth low-frequency cepstral coefficient, the proportion of descriptions of somatic symptoms, the ratio of the spectral index to the flow, descriptive structures, articulation dispersion, and other parameters. As a result of classification the CatBoost algorithm showed a weighted F1 measure of 76.16% and an accuracy of 75.64% which corresponds to the stated non-functional requirements. To balance rare classes, weights inverse to the number of examples were used, with a small constant added for stability, which were taken into account when calculating the loss function based on cross-entropy. AdamW was used as the optimizer,

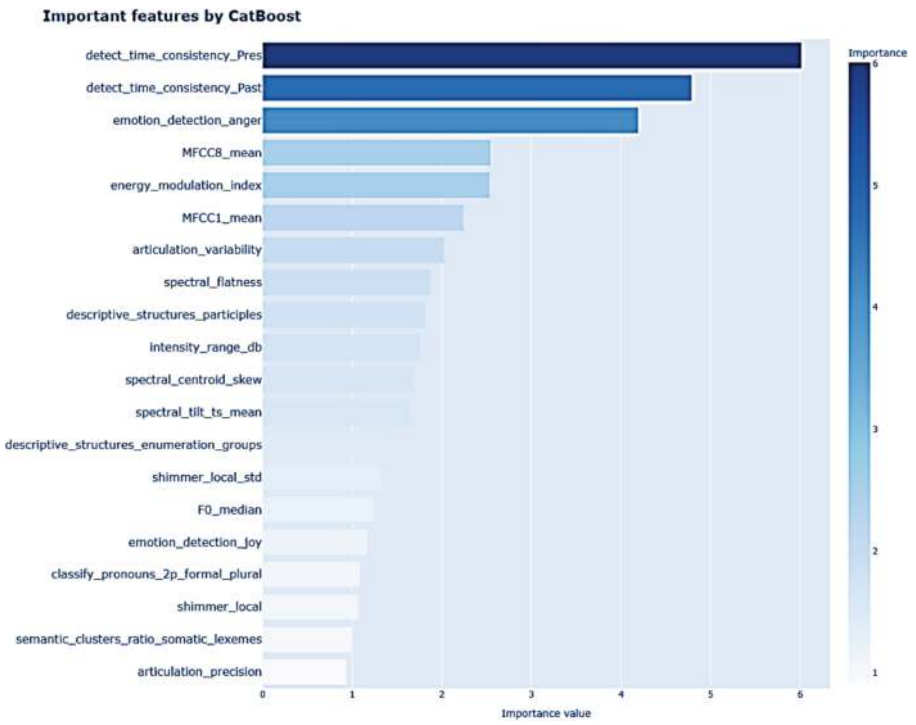


Fig. 5. Visualization of the 20 most important features (CatBoost).

and training was conducted for 60 epochs at a rate of 0.002 and a batch size of 128. The best F1 measure value of 62,2% was achieved in the 26 epoch, after which the model was saved for inference. The final results were a weighted F1 score of 62,2% and an accuracy of 73.6%. Analysis of the error matrix showed that the MFCC-based model is inferior in overall quality to the CatBoost classifier, but is more effective at identifying patients with depression. The graph of the loss function change is shown in Fig. 6.

The classification results are also presented in the form of a similarity matrix in Fig. 7.

The model was unable to identify consistent patterns based on MFCC and is crucial inferior in quality to the CatBoost classifier based on interpretable features, but it performs better in classifying patients with depression.

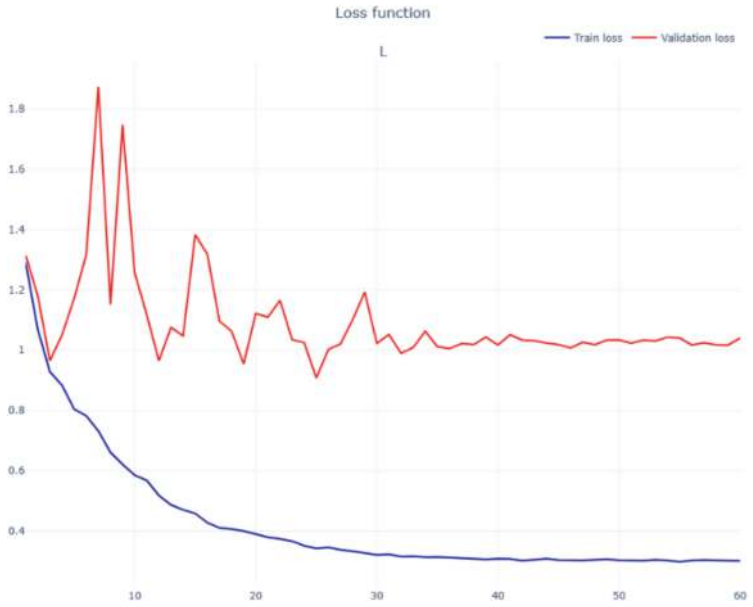


Fig. 6. Loss function for CNN.

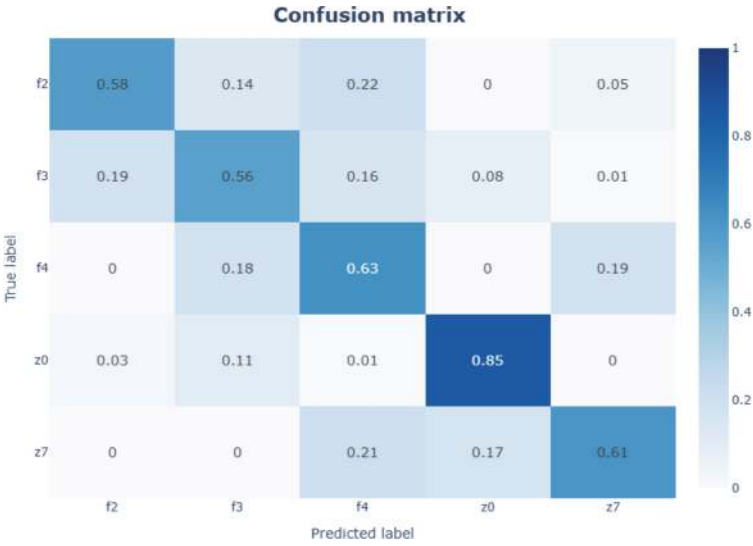


Fig. 7. Confusion matrix for evaluating the recognition quality of each class (CNN).

3 Conclusions

Developing a software product that is accessible to a wide range of users and capable of monitoring mental health with the ability to respond promptly to symptoms of physical and mental discomfort is a complex interdisciplinary task requiring the participation of

psychiatrists, clinical psychologists, clinicians, analysts, and developers. The increase in mental disorders is already having a negative impact on personal well-being, the emotional atmosphere in the workplace and family, and is also creating a significant economic burden. As part of the work, studies were conducted to update statistics on mental disorders and emotional states of the population and assess their economic impact on the Russian and global economies. A meta-analysis was conducted, combining data on studies of acoustic and linguistic markers, which formed the basis for modules for extracting interpretable features. A training dataset was formed, which became the basis for further development. As a result, two classifiers were created, the best of which showed a weighted F1 score of 76.16%. The accuracy of the models is expected to improve by increasing the consistency of experts in data collection, improving the quality of audio file preprocessing, and implementing more modern machine learning architectures.

References

1. Trofimov, D., Zapriy, M., Khomenko, A., Sloeva, E., Kotilevets, I., Smirnova, D.: Mapping affective profiles in depression, burnout, normal sadness, and euthymic state: a self-report screening tool developed through a machine learning approach. Psychiat. Danubina **37**(Suppl. 1), 237–259 (2025)
2. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Commun. **71**, 10–49 (2015). https://doi.org/10.1016/j.specom.2015.03.004
3. Prevalence of mental disorders [Electronic resource].: WHO Regional Office for Europe. https://gateway.euro.who.int/ru/indicators/hfa_391-2410-prevalence-of-mentaldisorders/#id=19322 Last accessed 14 Sep 2025
4. Anfinson, T.J., Kathol, R.G.: Screening laboratory evaluation in psychiatric patients: a review. Gen. Hosp. Psych. **14**(4), 248–257 (1992). https://doi.org/10.1016/0163-8343(92)90095-R
5. Tests and Assessments Used for Diagnosing Mental Health Disorders [Electronic resource].: Desert Hope Treatment Center. https://deserthopetreatment.com/co-occurring-disorders/tests-and-assessments Last accessed 15 Feb 2025
6. Shcherbakova, I.V.: Methods for diagnosing mental illnesses [Electronic resource]. Alvian Psychotherapy Center. https://medalvian.ru/zhurnal/metody-diagnostiki-psihicheskih-zabolevanij Last accessed 15 Feb 2025
7. Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., et al.: Tradeoffs between machine learning and deep learning for mental-illness detection on social media. Sci. Rep. **15**, 14497 (2025). https://doi.org/10.1038/s41598-025-99167-6
8. Nova, K.: Machine learning approaches for automated mental disorder classification based on social media textual data. Contemp. Issues Behav. Soc. Sci. **7**(1), 70–83 (2023)
9. Der Einfluss Digitaler Technologien AUF Moderne Unternehmensstrategien.: Zentralasiatische Zeitschrift für Innovationen in Wirtschaft und Sozialwissenschaften (CAJIBSS) [Electronic resource]. Zugriff. https://researchberg.com/index.php/CIBSS/article/view/149 Last accessed 12 Apr 2025
10. Early EEG data, machine learning can help predict antidepressant treatment response [Electronic resource]. The Times of India. https://timesofindia.indiatimes.com/city/chennai/early-eeg-data-machine-learning129can-help-predict-antidepressant-treatment-response-finds-study-by-scientists-fromiit-madras-and-czech-academy-of-sciences/articleshow/121297466.cms (2025). Last accessed 20 May 2025

11. Naderi, H., Haji Soleimani, B., Matwin, S.: Multimodal deep learning for mental-disorders prediction from audio speech samples. arXiv 1909.01067 (2020). https://doi.org/10.48550/arXiv.1909.01067
12. Sun, J., Liao, R., Shalaginov, M.Y., Zeng, T.H.: A machine-learning approach for predicting depression through demographic and socioeconomic features. medRxiv 2022.11.13.22281677 (2022). https://doi.org/10.1101/2022.11.13.22281677
13. Duangchaemkarn, K., Khammarew, P., Aramvith, S.: Machine learning-based classification of mental health state using the DASS-21 profile. In: Proceedings of the 2024 Biomedical Engineering International Conference (BMEiCON) (2024). https://doi.org/10.1109/BMEiCON64021.2024.10896316
14. Kwon, Y., Blazyte, A., Jeon, Y., et al.: Identification of 17 novel epigenetic biomarkers associated with anxiety disorders using differential methylation analysis followed by machine-learning-based validation. Clin. Epigenet. **17**, 24 (2025). https://doi.org/10.1186/s13148-025-01819-x
15. Bracher-Smith, M., Crawford, K., Escott-Price, V.: Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol. Psychiatry **26**, 70–79 (2021). https://doi.org/10.1038/s41380-020-0825-2
16. Lundqvist, D.: The Karolinska directed emotional faces—KDEF. In: Lundqvist, D., Flykt, A., Öhman, A. (eds.) Karolinska Institutet: [Electronic resource]. https://kdef.se (1998). Last accessed 04 May 2025
17. Viera, A.J.: Understanding interobserver agreement: the kappa statistic. In: Viera, A.J., Garrett, J.M. (eds.) Family Medicine, vol. 5, pp. 360–363. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052 (2005). Last accessed:18 May 2025
18. Chi-Square Test of Independence.: JMP Statistics Knowledge Portal: [Electronic resource]. https://www.jmp.com/en_us/statistics-knowledge-portal/chi-squaretest/chi-square-test-of-independence.html. Last accessed 18 May 2025
19. Cramér's, V.: IBM Documentation: [Electronic resource]. https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-cramrs-v. Last accessed 18 May 2025
20. Kullback-Leibler Divergence Explained.: Count Bayesie Blog: [Electronic resource]. https://www.countbayesie.com/blog/2017/5/9/kullback-leiblerdivergence-explained (2017). Last accessed 18 May 2025

Conceptual Model of Value Chain Management of a Digital University Based on Artificial Intelligence

Alexander Usanov[1(✉)], Oksana Moshchenko[1], and Vinodha D[2]

[1] Financial University Under the Government of the Russian Federation, Moscow, Russia
ayusanov@fa.ru
[2] Christ University, Bangalore, India

Abstract. This paper addresses the challenge of transforming the management system of the contemporary university amidst widespread digitalization. The study's relevance stems from the imperative to shift from conventional methods to data-driven and AI-driven governance models. The primary goal is to develop a conceptual model for managing the university value chain leveraging artificial intelligence (AI) technologies. The research methodology is grounded in a systems approach, Michael Porter's value chain concept, and digital transformation theory. The principal outcome is a proprietary Digital University Value Chain (DUVC) model, which positions AI as a pervasive intelligent layer that integrates every stage of educational, research, and social value creation. The efficacy of this model is substantiated through a case study of the Financial University under the Government of the Russian Federation. It is concluded that adopting an AI-driven approach is a strategic determinant for enhancing a university's competitive advantage.

Keywords: Artificial intelligence · Digital university · Value chain · Digital transformation · Higher education management · AI-driven model · Data analysis · Management model

1 Introduction

The current stage of higher education development is marked by a paradigmatic shift driven by the imperatives of the digital economy, globalization processes, and the accelerated infiltration of artificial intelligence (AI) technologies into academic and managerial Practice. Universities, Acting as Key Actors in National Innovation Ecosystems, Are transitioning from a traditional, function-oriented model to the concept of the Digital University. This concept implies the creation of a data-driven organization that relies on predictive analytics and the intelligent automation of core processes. These transformations are not local but systemic, affecting the strategic and organizational foundations of university functioning and requiring a fundamental rethinking of approaches to managing the value creation chain.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 685–695, 2027.
https://doi.org/10.1007/978-3-032-27834-0_60

The concept of the value chain, introduced by M. Porter as a tool for strategic analysis in commercial organizations [1], has proven its universality and has been successfully adapted for the analysis of organizations in the public and non-profit sectors. In the context of higher education, this methodology allows the decomposition of a university's activity as an integrated system in which input resources (academic capital, applicants, technologies, financial resources) are transformed into final outputs (competent graduates, scientific discoveries, innovative products) that generate measurable socio-economic value for society. However, existing models of university management often lack a holistic approach to the measurement, analysis, and optimization of this value using artificial intelligence tools.

Recent studies in the fields of management and digital transformation [2–4] empirically confirm that the digital transformation of educational organizations requires not only technological modernization but also the formation of a new managerial paradigm based on data-driven and AI-driven approaches. Artificial intelligence is evolving from a tool for isolated automation into a strategic asset that enables comprehensive analysis of large volumes of data concerning the student body, educational processes, research performance, and resource efficiency. The implementation of AI technologies contributes to enhancing the transparency and validity of managerial decisions, creating personalized educational trajectories, optimizing interactions with external stakeholders, and, consequently, increasing the competitiveness of universities in the global knowledge economy.

Thus, the relevance of the present study stems from the objective need to develop a conceptual model for managing the university value chain that integrates artificial intelligence technologies. Such a model is designed to unite the strategic, organizational, and technological dimensions of university digital transformation, providing a holistic and proactive framework for managing the creation of educational, research, and social value.

The scientific novelty and practical significance of this research are closely interrelated and lie in the development of an integrated conceptual model for managing the university value chain based on the end-to-end integration of artificial intelligence technologies. The scientific novelty of the work primarily consists in the author's proposed *Digital University Value Chain (DUVC)* model, which systematically combines M. Porter's classical concept with the modern AI-driven management paradigm, operationalizing university activities through stages of value creation and identifying the key points of AI application. Furthermore, the study introduces, for the first time, the concept of AI as a *"pervasive intelligent layer"* within the digital university architecture—serving as a systemic mechanism that ensures the analytical coherence and predictive capability of all processes. This approach provides a theoretical justification for the necessary shift from a retrospective data-driven paradigm to an AI-driven model focused on scenario modeling and proactive decision-making. For quantitative assessment, a system of balanced indicators and an integrated *Digital Value Index (DVI)* are proposed, forming a methodological foundation for measuring the effectiveness of digital initiatives.

These scientific results directly determine the practical significance of the study for university administrators and vice-rectors responsible for digital transformation. The developed model can serve as a strategic "roadmap" for designing and implementing

digital development programs, enabling a transition from fragmented automation to systemic process reengineering. The proposed methodology and system of indicators can be used to conduct digital maturity audits, identify bottlenecks, and substantiate investments in specific AI projects. The analysis of empirical cases from the Financial University and Lomonosov Moscow State University is particularly valuable, as it demonstrates two different yet successful strategies for AI integration, allowing other universities to choose the most relevant path. Finally, the managerial recommendations formulated regarding the establishment of competence centers, the implementation of AI governance, and the development of staff digital skills can be directly applied to university management practice, providing a comprehensive foundation for the strategic transformation of higher education institutions in the digital economy.

2 Theoretical and Methodological Foundations of the Study

The theoretical foundation of this research is based on M. Porter's concept of the value chain, which views an organization as a set of interrelated activities aimed at creating customer value [1]. Within this framework, primary and supporting activities are distinguished. When applied to the higher education sector, the adaptation of this logic implies the following interpretation of the key components: attraction, selection, and adaptation of applicants; design and implementation of educational programs; organization and support of research activities; knowledge and technology transfer; career support for graduates and interaction with the labor market. Supporting activities include human capital management, IT infrastructure development, financial and economic management, administrative governance, and the development of partnership networks.

The modern scientific discourse expands the classical interpretation of value, shifting the focus toward the concept of *value co-creation*, which involves the active participation of students, faculty, employers, and regional communities [5, 6]. This approach highlights the evolution of the university's role from that of an educational service provider to a coordinator of a complex knowledge ecosystem, where value is generated through continuous interaction among all participants.

Digital Transformation (DT) in education is interpreted not as mere automation or digitization of existing processes but as a comprehensive restructuring of the business model, organizational structure, and corporate culture based on data and digital technologies [2, 3]. To assess the depth and systemic nature of these changes, *Digital Maturity Models (DMMs)* are applied. These models, developed by leading consulting and research centers [7–9], emphasize the multidimensionality of transformation, including such aspects as digital educational services, research analytics, student experience management, operational efficiency, and data governance maturity. Post-pandemic analyses record an accelerated diffusion of technological solutions (LMS/EMS platforms, cloud infrastructure, proctoring systems, learning analytics); however, they also reveal systemic challenges—fragmented digitalization and weak integration of solutions at the level of institutional data architecture [7, 10].

The use of AI in higher education is developing along two interconnected trajectories:

(1) *Pedagogy-centric AI*, which includes adaptive learning platforms, intelligent tutoring systems, automated assessment, and generative tools for academic writing support; and
(2) *Management-centric AI*, which encompasses predictive analytics for student dropout and performance risks, optimization of timetables and classroom workloads, data-driven budgeting models, and research productivity monitoring [11, 12]. Studies indicate that the greatest synergistic effect is achieved through the end-to-end, seamless integration of AI along the entire value chain—from applicant orientation to graduate career trajectory analysis. In the management domain, AI enhances forecasting accuracy, reduces transaction costs, and enables real-time scenario modeling, which is critically important under conditions of demographic turbulence and intensified global competition for talent [13].

The key barriers to scaling AI in the educational environment are related to data quality, interoperability, and ethical considerations. International organizations [7] emphasize the need to:

(a) build a robust *Data Governance* system within universities, including data catalogs, defined wnership roles, and the implementation of quality and metadata standards;
(b) ensure transparency and interpretability of algorithmic decisions (*Explainable AI, XAI*);
(c) guarantee the protection of students' personal data; and
(d) prevent algorithmic discrimination and bias.

In the context of *learning analytics*, particular importance is given to obtaining informed consent and ensuring students' right to an explanation of decisions made based on their digital footprint [14].

The academic literature unanimously stresses that the effectiveness of AI initiatives depends less on the set of technologies used and more on the presence of *dynamic capabilities* within the organization: digital leadership, a mature digital culture, relevant staff competencies, effective mechanisms of cross-functional coordination, and a robust *Enterprise Architecture* [2, 4]. To achieve sustainable impact, it is essential to transition toward a platform-based architecture, standardize APIs for data exchange, implement *MLOps* and *DataOps* practices, and launch continuous improvement cycles for educational and administrative processes. Empirical evidence shows that in the absence of these conditions, AI remains at the level of local innovations and fails to translate into measurable value growth for stakeholders [15, 16].

The conducted literature analysis allows identifying several scientific problems that remain insufficiently addressed:

1. **The absence of comprehensive models** describing an end-to-end *AI-driven* value chain with a system of metrics that organically integrates academic, financial–economic, and social outcomes [17].
2. **Insufficient elaboration of mechanisms** for integrating ethical principles and AI trust requirements into the corporate governance system of universities.
3. **A lack of empirical studies** demonstrating causal relationships between the implementation of AI solutions and key university performance indicators (student retention, graduate employability, cost structure).

4. **The absence of reference architectural patterns** for scaling AI within the university environment, including role models, *MLOps* processes, and data catalogs.
5. **Limited exploration of mechanisms** that connect the concept of *co-creation* with algorithmic decision-making systems [18].

Thus, the contemporary research agenda is shifting from the study of isolated digital solutions toward the design of a holistic *AI-driven* model for managing the university value chain. The development and operationalization of such a model through a system of measurable indicators constitute the principal scientific contribution of this study.

The methodological framework of the study is based on an interdisciplinary synthesis of the value chain concept [1], the theory of digital transformation [18, 19], and data-driven management and AI application approaches in education [20]. A system approach was employed, within which the university is viewed as an open socio-economic system, where digital technologies and AI act not only as optimization tools but also as structural elements of the organizational architecture [21, 22].

For operationalization purposes, the university value chain is decomposed into the following interrelated blocks:

- **Inputs**: Resource base—applicants, faculty members, infrastructure, data, financial capital, and institutional reputation.
- **Core Activities**: Core operations—educational activities (program design and implementation), scientific research, innovation projects and technology transfer, and collaboration with industrial partners and government bodies.
- **Outputs**: Direct results—graduates with defined competencies, scientific publications, patents, grants, and newly created startups.
- **Outcomes**: Long-term impacts—social, economic, and intellectual value for regions, the national economy, and society as a whole [23].

Each element is characterized by a set of digital metrics and represents a potential application point for AI technologies to enhance efficiency. Collectively, these elements form the AI-driven Value Chain Model (AIVC).

The study was implemented in three sequential stages:

1. **Diagnostic Stage**: A comparative analysis of digital maturity and AI integration levels in managerial processes was conducted for leading universities in Russia and worldwide (including the Financial University under the Government of the Russian Federation, Lomonosov Moscow State University, MIT, and the Open University UK) to identify best practices and barriers.
2. **Modeling Stage**: A conceptual model of the digital value chain was developed, defining the application points, functions, and expected effects of AI utilization at each stage.
3. **Analytical and Evaluation Stage**: The effectiveness of traditional and AI-driven models was compared across key performance indicators, digital maturity, and management system adaptability.

Content Analysis: Strategic documents were examined, including digital transformation programs, university development strategies, annual reports, and open data for the period 2020–2025.

Expert Evaluation (Delphi Method): A two-round survey was conducted involving 15 experts in digital education, university management, and AI technologies to validate the model and verify the indicator system.

Digital Maturity Indicators: A modified Gartner Digital Maturity Assessment model was applied, incorporating five levels of maturity: Initial—Developing—Integrated—Optimized—AI-driven.

Visualization and Analytical Methods: The value chain diagram and AI integration matrix were constructed using Business Intelligence platforms (Power BI) and Python-based analytical scripts.

For the quantitative assessment of the effects of AI implementation, a system of balanced indicators was proposed, grouped into four categories (Table 1).

Table 1. System of indicators for quantitative assessment of AI implementation effects.

Group	Examples of indicators	Data sources
Educational	Student retention rate, employability index, satisfaction index	LMS, CRM systems, surveys
Research and innovation	Number of Q1–Q2 publications, volume of attracted grants, citation index, number of patents	*Scopus*, *RSCI*, internal reports
Financial and managerial	Return on investment (ROI) in digital projects, cost efficiency, decision-making speed	*ERP*, *Power BI*
Social and reputational	Alumni engagement index, positions in digital rankings, social impact assessment	External rankings, social media analytics

Based on this system, an integral Digital Value Index (DVI) of the university is formed, calculated using a weighted formula.

3 Research Results

The outcome of the research was the construction of a conceptual model of the Digital University Value Chain (DUVC), systematically integrated with artificial intelligence technologies. The model reflects a closed value creation cycle and includes four key blocks: inputs, core processes, outputs, and outcomes, where AI performs the role of a central analytical core.

The analysis showed that AI functions as a "pervasive intelligent layer," ensuring seamless integration of data and analytics between all elements of the chain (Table 2).

Thus, artificial intelligence becomes the "digital nervous system" of the university, ensuring data integration, forecasting, and decision support across all managerial and academic processes. A comparative analysis of the traditional and AI-driven management models is presented in Table 3.

Table 2. Conceptual model of the digital university value chain (DUVC).

Value chain stage	AI functions	Key effects
Admissions and career guidance	Predictive analytics of applicants, scoring models, NLP analysis of motivation letters	Increased applicant conversion rate, reduced marketing costs
Educational process	Adaptive learning trajectories, intelligent tutors, chatbots, generative assistants	Personalized learning, enhanced student engagement, reduced dropout rates
Research activity	Automated data analysis, hypothesis generation, pattern recognition, recommendation systems	Growth of publication activity, optimization of grant applications
Management and finance	Predictive analytics, AI-based budgeting, intelligent dashboards, scenario modeling	Improved cost efficiency, acceleration of managerial decision cycles
Partnership and external engagement	Labor market analysis, ML models for employer matching, partner network analytics	Increased graduate employability, strengthening of institutional reputation capital

Table 3. Comparative analysis of traditional and AI-driven management models.

Indicator	Traditional model	AI-driven model
Decision-making	Based on experience, intuition, and retrospective data	Based on data, models, and scenario analysis
Learning management	Standardized educational programs	Personalized learning trajectories and adaptive recommendations
Quality control	Periodic, post-factum	Continuous, predictive, real-time
Financial management	Reactive, based on approved budgets	Proactive, optimization-oriented, scenario-based
Stakeholder interaction	One-way, informational	Interactive, with elements of co-creation

Empirical analysis has demonstrated that the transition to an AI-driven model is accompanied by a 15–25% increase in operational efficiency, a reduction in transactional costs of up to 20%, and an improvement in the accuracy of student retention forecasts by up to 30%.

Empirical Case Study: Financial University under the Government of the Russian Federation

To validate the model, an in-depth case study of the Financial University was conducted. Between 2020 and 2024, the university implemented a series of projects aimed at intellectualizing its management processes. Key examples include:

- Educational Data Analytics: An LMS module for analyzing students' digital footprints was introduced, enabling high-precision forecasting of dropout risks.
- AI Assistants: Generative models (e.g., ChatGPT, Copilot) were actively used for preparing educational materials and automating routine tasks.
- Power BI Dashboard: A unified management analytics system was created, integrating data from various systems in real-time.
- AI-driven Assessment of Scientific Activity: Algorithms were developed for the objective ranking of research productivity.

Implementation Results (2020–2024): an increase in the accuracy of student academic performance forecasts (AUC from 0.68 to 0.91); a 4.5-fold reduction in the time required to prepare analytical reports; and a 27% growth in publications in Scopus and WoS. The experience of the Financial University confirms that AI can become the core of the digital value chain, ensuring a sustainable increase in efficiency and quality.

Empirical Case Study: Lomonosov Moscow State University (MSU)

As a contrasting example demonstrating a different strategy for AI integration, the experience of Lomonosov Moscow State University was analyzed. While the primary driver of transformation at the Financial University is enhancing managerial efficiency, MSU places a greater emphasis on the intellectualization of fundamental scientific research and the support of complex educational programs. The transformation at MSU is less centralized and is largely initiated at the faculty and research center level, which possess unique competencies and resources.

Key examples of AI and digital technology integration include:

- **Intellectualization of Scientific Research via a Supercomputing Complex**: The "Lomonosov" and "Lomonosov-2" supercomputers form the core for applying machine learning and AI to resource-intensive tasks. Algorithms are used for modeling complex physical processes, analyzing genomic data in bioinformatics, developing new materials with predefined properties, and climate modeling. Here, AI functions not as a managerial tool but as a scientific instrument that directly creates research value.
- **Development of Digital Educational Platforms**: MSU is actively developing its own online education platforms, such as "University without Borders." These platforms incorporate AI elements to create course recommendation systems, adaptive testing systems that adjust task difficulty to the student's level, and automated analysis of essays and written assignments for plagiarism and style.
- **Intelligent Analysis and Search Systems for Digital Archives**: Given the vast volume of accumulated scientific and historical data, MSU is developing intelligent search systems. Using Natural Language Processing (NLP) and computer vision technologies, digitized manuscripts, rare books, and scientific reports are indexed and semantically analyzed, allowing researchers to uncover implicit connections within large arrays of unstructured data.
- **Predictive Analytics in Student Body Management**: Several faculties (notably, the Faculty of Computational Mathematics and Cybernetics and the Faculty of Mechanics and Mathematics) are piloting systems to analyze student performance for the early detection of academic failure risks. Models based on data from the LMS and

electronic grade books enable curators and instructors to provide timely support to underperforming students.

Implementation Results (assessment for the 2020–2024 period):

- Acceleration of computational experiment cycles in breakthrough scientific fields (estimated at 3–5 times), leading to an increase in publications in high-ranking international journals.
- A more than 40% growth in the audience reach of online courses and increased student engagement through the implementation of adaptive technologies.
- Up to a 70% reduction in the time required for researchers to find relevant information in scientific collections and archives using the new intelligent systems.
- Improved accuracy in predicting academic risks at pilot faculties, which contributed to a 5–7% reduction in dropout rates within target groups.

Thus, the MSU case study demonstrates a digital transformation model where AI is primarily integrated into core activities—science and education. Managerial effects here are a consequence of increased efficiency in key processes rather than an end in themselves. This approach is characteristic of major research universities where the creation of new knowledge is the main link in the value chain.

4 Discussion of Results

The research findings confirm that the digital transformation of universities should be viewed as a strategic transition to a new value chain management model, with AI as its central element. A comparison of the models shows that AI radically changes the architecture of university governance, shifting it from reliance on expert assessments to predictive and cognitive analytics. The value chain is transformed from a static structure into a dynamic, adaptive ecosystem.

These conclusions align with the works of Westerman (2014) and the OECD (2022), which indicate that the systemic integration of AI and data analytics is the primary success factor in digital transformation. However, while leading Western universities demonstrate a high level of maturity, Russian universities are predominantly in the integration stage, characterized by localized AI initiatives but lacking a unified data architecture and end-to-end governance.

5 Conclusions

The need to create competence centers for artificial intelligence and data analytics stems from the growing role of digital technologies in managing educational and scientific processes. Such centers will become the core of methodological and technological support for the implementation of AI solutions, ensuring the consolidation of expertise, the development of human capital, and the transfer of best practices. An important direction is the formation of a corporate AI Governance policy, which includes ethical standards for the use of artificial intelligence and principles of algorithm explainability. This will increase trust in digital solutions and minimize the risks of their incorrect application.

Concurrently, it is necessary to integrate AI metrics into the KPI system for managers and faculty, which will link digital transformation to the performance evaluation system for managerial and pedagogical decisions. The final element is the development of digital competencies among all participants in the educational process, fostering a culture of conscious and responsible use of artificial intelligence in the educational environment.

The risks and limitations of implementing artificial intelligence in the educational environment are driven by both technological and institutional factors. One of the key challenges involves ethical risks and the potential for algorithmic discrimination that arises from the use of automated decision-making systems and analytical models. A significant threat is the digital divide between universities, related to disparities in digital infrastructure, access to computing resources, and personnel qualifications. A further limitation is the insufficient interoperability of legacy IT systems, which complicates the integration of new digital solutions and leads to a fragmented information environment. Finally, a significant barrier to digital transformation remains the resistance of personnel to organizational changes, caused by both a lack of digital competencies and the institutional inertia of management structures.

The conducted research has allowed for the development and validation of a conceptual model of the Digital University Value Chain (DUVC), in which artificial intelligence acts as a systemic mechanism for intelligent management, ensuring end-to-end analytical connectivity, predictability, and process adaptability. The empirical analysis confirmed that the application of AI in university management leads to a significant increase in institutional efficiency.

For the effective implementation of artificial intelligence into the university management system, it is advisable to institutionalize AI governance processes, providing regulatory, ethical, and organizational support for the use of intelligent technologies. It is necessary to form a unified corporate data architecture, which will become the foundation for end-to-end analytics, the integration of information flows, and increased transparency of managerial decisions. It is important to embed AI tools into the strategic and financial management system, which will improve forecasting accuracy, optimize resource allocation, and enhance the performance of the educational organization. A key direction is also the cultivation of a data-driven and evidence-based management culture, focused on making decisions based on objective data and empirical evidence.

References

1. Porter, M.E.: Competitive advantage: creating and sustaining superior performance. Porter, M.E., New York: Free Press, 557 p. ISBN 978-0684841465 (1985)
2. Kane, G.C., Palmer, D., Phillips, A.N., et al.: Coming of Age Digitally. MIT Sloan Management Review and Deloitte Insights. https://sloanreview.mit.edu/projects/coming-of-age-digitally/ (2019)
3. Westerman, G.: Leading digital: turning technology into business transformation. Westerman, G., Bonnet, D., McAfee A., Boston: Harvard Business Review Press, 304 p. ISBN 978-1625272478 (2014)
4. AI and education: guidance for policy-makers. Paris: UNESCO, 45 p. https://unesdoc.unesco.org/ark:/48223/pf0000385755 (2023)

5. Prahalad, C.K.: The future of competition: co-creating unique value with customers. Prahalad, C.K., Ramaswamy V., Boston: Harvard Business School Press, 257 p. ISBN 978-1578519538 (2004)
6. Aoun, J.E.: Robot-proof: higher education in the age of artificial intelligence. J. E. Aoun.— Cambridge: The MIT Press, 216 p. ISBN 978-0262037282 (2017)
7. Education at a Glance 2022: OECD Indicators. Paris: OECD Publishing, 456 p (2022). https://doi.org/10.1787/3197152b-en
8. Hype Cycle for Higher Education: Gartner, Inc. https://www.gartner.com/en/industries/higher-education (2023)
9. Digital maturity framework for higher education: Jisc. https://www.jisc.ac.uk/guides/digital-maturity-framework-for-higher-education (2022)
10. Selwyn, N.: Is technology good for education? Selwyn N., Cambridge: Polity Press, 224 p. ISBN 978-0745688536 (2016)
11. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int. J. Educ. Technol. High. Educ. **16**(1) (2019). https://doi.org/10.1186/s41239-019-0171-0
12. Ifenthaler, D.: Utilising learning analytics to support study success in higher education. Ifenthaler, D., Yau J.Y.K., J. Comput. Educ. **7**(1), 97–116 (2020). https://doi.org/10.1007/s40692-019-00150-3
13. Brynjolfsson, E.: The second machine age: work, progress, and prosperity in a time of brilliant technologies. Brynjolfsson, E., McAfee A., New York: W.W. Norton & Company, 320 p. ISBN 978-0393239355 (2014)
14. Siemens, G.: Penetrating the fog: analytics in learning and education. Siemens, G., Long, P., EDUCAUSE review **46**(5), 30–32. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education (2011)
15. Christensen, C.M., Eyring, H.J.: The innovative university: changing the DNA of higher education from the inside out. Jossey-Bass (2011)
16. Davenport, T.H., Harris, J.G.: Competing on analytics: the new science of winning. Harvard Business Review Press (2007)
17. Schwab, K.: The fourth industrial revolution. Crown Business (2016)
18. Fullan, M.: The new meaning of educational change (5th ed.). Teachers College Press (2015)
19. Tapscott, D.: The digital economy: rethinking promise and peril in the age of networked intelligence. McGraw-Hill (1996)
20. EDUCAUSE: (annual editions). The horizon report: higher education edition. https://www.educause.edu/horizon-report
21. Selingo, J.J.: College (un)bound: The future of higher education and what it means for students. Houghton Mifflin Harcourt (2013)
22. Barber, M., Donnelly, K., Rizvi, S.: An avalanche is coming: higher education and the revolution ahead. Institute for Public Policy Research (2013)
23. Picciano, A.G.: Online education: foundations, planning, and pedagogy. Routledge (2017)

Developing an Information Security Risk Assessment Model Using the CVSS

Maria Utenkova[1(✉)], Andrey Komkov[1], Elena Maksimova[2], and Oksana Mezentseva[3]

[1] MIREA – Russian Technological University, Moscow 119048, Russian Federation
utenkova@mirea.ru
[2] Moscow Technical University of Communications and Informatics, Moscow 111024, Russian Federation
[3] North Caucasus Federal University, Stavropol, Russia

Abstract. As digitalization continues to become an increasingly prominent trend, organizations are facing the challenge of optimizing resource allocation while ensuring their cyber resilience. In this context, we have examined various risk management methods and techniques, identifying their benefits and drawbacks. To address some of these limitations, we propose an improved method for quantifying information security risks, taking into account the severity of vulnerabilities as assessed by Common Vulnerability Scoring System (CVSS). This approach includes an enhanced risk assessment model that considers the impact of vulnerabilities on overall risk. Special attention has been paid to the process of information security risk assessment, including the description of our method using linguistic fuzzy logic. To validate the effectiveness of our proposed risk assessment model, we conducted an experiment on a sample network architecture, resulting in the identification of the assets with the highest potential risk.

Keywords: Risk assessment · CVSS · Information security · Unacceptable events · Sources of threats · Linguistic fuzzy logic

1 Introduction

The projects "Digital Economy" and "Data Economy and Digital Transformation of the State" [1] pay special attention to the federal project "Cybersecurity Infrastructure". This project aims to ensure the security of the Russian Internet, counteract online fraud, and prevent personal data leaks. Achieving these objectives will enhance the cyber resilience of organizations [2], contributing to the stability of the national economy [3].

It is worth noting that industry 4.0, which involves the adoption of innovative technologies in production processes, also raises concerns regarding information security [4]. This highlights the importance of addressing cybersecurity issues to ensure a stable and secure digital environment.

Increasing cyber resilience is possible through increasing awareness among employees, customers, and partners of organizations [5, 6]. However, to increase the effectiveness of information protection tools for digital systems, it is important to assess

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 696–707, 2027.
https://doi.org/10.1007/978-3-032-27834-0_61

information security risks [7]. This step allows us to obtain information about current threats to information security and helps to allocate resources more optimally. This is particularly relevant for small and medium-sized businesses with limited budgets for information security [8]. Information security risk assessment models can help identify threats that could cause significant losses if exploited. In some cases, these models can even predict future events [9, 10]. It is important to consider that many threat scenarios, such as privilege abuse, identity theft, and data leakage, can occur during insider attacks [11]. It is also worth noting that risks in supply chains should be taken into account, as attacks on suppliers (e.g., providers) can serve as a starting point for attacks on the target organization [12].

Currently, Zhukov et al. [13], McNett [14], Renney et al. [15], Achaal et al. [16], Fernandez et al. [17], Alva et al. [18] and Ao et al. [19] are all working on this topic, but the solutions they propose mainly use non-universal models for vulnerability assessment. In these approaches, decisions on parameter values are based on expert opinions and complex mathematical models. This is why we developed our own model to simplify threat assessment.

A weak information security organization is dangerous, as it leads to financial, reputational, and legal damage [20–23]. This makes information security issues critically important at the moment.

The novelty of our approach lies in combining linguistic fuzzy logic with quantitative vulnerability risk assessment and a basic methodology for information security threat assessment.

The aim of this research is to develop a model for evaluating information security risks by identifying the CVSS (Common Vulnerability Scoring System) information security vulnerability factor in order to enhance the efficiency of information security measures employed.

To achieve this goal, it is necessary to complete the following tasks:

1. Conduct an analysis of the subject area.
2. Develop an information security risk assessment model using the Common Vulnerability Scoring System (CVSS).
3. Test the information security risk assessment model experimentally.

2 Analysis of the subject area

2.1 Analysis of Information Security Risk Assessment Methods and Methodologies

To date, there are several methods and methodologies available for assessing information security risks. Some of the most commonly used include ISO 27005, FRAP, STO BR IBBS +716-P, OCTAVE and FAIR. These are compared in Table 1.

Each unique situation may require its own methodology. What may be seen as a disadvantage for one organization could turn out to be a significant advantage for another.

Table 1. A comparison of the advantages and disadvantages of different information security risk assessment methods.

	Advantages	Disadvantages
ISO 27005	It has no restrictions on the size of organizations or their fields of activity. There are various scales and methods for assessing risks	Highly qualified specialists are required. It is also necessary to update GOST R ISO/IEC 27005–2010
FRAP	Applicable to various organizations and simplifies the information security risk assessment process, eliminating the need for highly qualified specialists	It does not provide recommendations on risk management and requires adaptation to the regulatory legal acts of the Russian Federation
STO BR IBBS +716-P	It provides examples of both risk assessment and risk management techniques	It is linked to financial industry organizations and requires highly qualified employees
OCTAVE	Applicable to various organizations and does not require high costs or the involvement of a third party	Highly qualified specialists are required. It requires adaptation to the regulatory legal acts of the Russian Federation
FAIR	It includes a detailed classification of risk factors and their interactions. It also provides a high-quality assessment of the likelihood of risk occurrence and the potential amount of losses	It is aimed at working with large organizations and requires high costs and a team of highly qualified specialists. It requires adaptation to the regulatory legal acts of the Russian Federation

2.2 Analysis of the CVSS Vulnerability Risk Assessment Method

CVSS is an open standard for quantifying vulnerability risks based on a set of parameters that can vary from version to version [24]. The latest version of CVSS, 4.0, includes the following basic metrics:

- Exploitability metrics: attack vector, attack complexity, attack requirements, privileges required, user interaction.
- Vulnerable system impact metrics: confidentiality, integrity, availability.
- Subsequent system impact metrics: confidentiality, integrity, availability.

In addition to the basic metrics, this standard also includes additional metrics that can change over time or depend on the specific context. These metrics are assigned a qualitative value, which is then used to calculate a quantitative assessment of vulnerability risk according to the algorithm outlined in the CVSS specification.

It's worth noting that previous versions of CVSS are still relevant and continue to be actively used. For example, vulnerabilities released in the last year in the database of FSTEC (Federal Service for Technical and Export Control) of Russia typically have

a CVSS 2.0 or 3.0 rating, although they may not have yet received a CVSS 4.0 rating. This is also true for foreign vulnerability databases [25].

Despite the fact that the range of possible assessment values for all versions of CVSS remains unchanged at [0,10], it's important to note that the assessment of a given vulnerability may vary depending on the specific circumstances.

2.3 Analysis of Software Tools Used in the Information Security Risk Assessment Process

Information security risk assessment can be conducted using various software tools. The requirements for these tools and their functionality depend on various factors, including the system being assessed and any additional restrictions that may apply [26, 27].

Within the framework of this study, several stages of information security risk assessment have been identified:

- Defining unacceptable events;
- Definition of assets;
- Identifying threat sources;
- Identify current threats and vulnerabilities;
- Prioritization of current threats.

Different software tools may be needed for each stage of information security risk assessment (see Table 2). It's worth noting that decision support systems play a crucial role in analyzing data collected during previous stages [28].

Table 2. Software tools used in various stages of information security risk assessment

Stage	Software tools
Defining unacceptable events	Intelligent expert management tools
	Assessment sheet management software
Definition of assets	Enterprise asset management systems
	Software for IT asset management
Identifying threat sources	A program for building a model of a violator
	Information security audit program
Identify current threats and vulnerabilities	Vulnerability scanner
	CVSS Assessment Program
	Automated application security testing tool
Prioritization of current threats	DSS for assessing destructive effects
	DSS for evaluating the effectiveness of protection systems
	A program for logical and probabilistic assessment of structural stability

Some organizations may require additional software tools that are not listed in this proposal. However, the software tools presented here are sufficient for the purposes of this study.

3 Develop an information security risk assessment model using the CVSS

3.1 A Formalized Presentation of the Information Security Risk Assessment Model

To analyze existing information security risks and predict their impact on the system, it is necessary to assess and prioritize risks, simulate possible threats, identify the most dangerous ones for the information system, and build necessary protection.

In accordance with the developed model (see Fig. 1), the following tasks need to be performed:

1a. Defining unacceptable events: It is essential to identify events that could cause significant disruptions to the organization's operations and become serious obstacles to achieving its operational and strategic objectives. For this purpose, it is necessary to work with the heads of key departments in the company to assess information security threats. At the end of this process, the main areas of protection should be identified and criteria for handling unacceptable events should be established. The purpose of establishing criteria for handling these events is to analyze the likelihood of them occurring within the organization's daily information technology infrastructure. This will help to ensure that appropriate measures are taken to prevent or minimize the impact of such events.

1b. Determining the budget for information security: They have limited resources to allocate to information security measures.

2. Definition of assets: In order to understand what is within the system and to evaluate these resources, finding the most important and valuable ones. At this stage, it is necessary to determine which assets require more reliable protection and who should have access to them. Systems and networks are also considered objects of protection, and a model of the organization's IT environment is created. Within this model, the components of the information system, as well as connections and interactions between them, are distinguished.
3. Identifying threat sources: At this point, it is essential to identify and categorize potential sources of threats to information security. Determine the types of attackers that could affect the system and classify each type as internal or external. Identify the purposes for which an attacker might carry out their actions. Describe the potential damage that could result from successful attacks and the objects that could be targeted by an attacker in order to influence the system.
4. Identify current threats and vulnerabilities. A list of potential threats that may affect the system being studied should be compiled. This includes an assessment of the likelihood of information security threats, as well as possible scenarios for how these threats could occur within the information system. It is important to create a

comprehensive list of these threats and describe how they could potentially attack the system through various vectors.

5. Prioritization of current threats. Threats are ranked based on a developed model. The most significant and investment-worthy vulnerabilities are identified, and these are addressed first. A list of threats that present the greatest risk and require immediate attention is compiled. Limited security resources are directed towards the most appropriate areas.

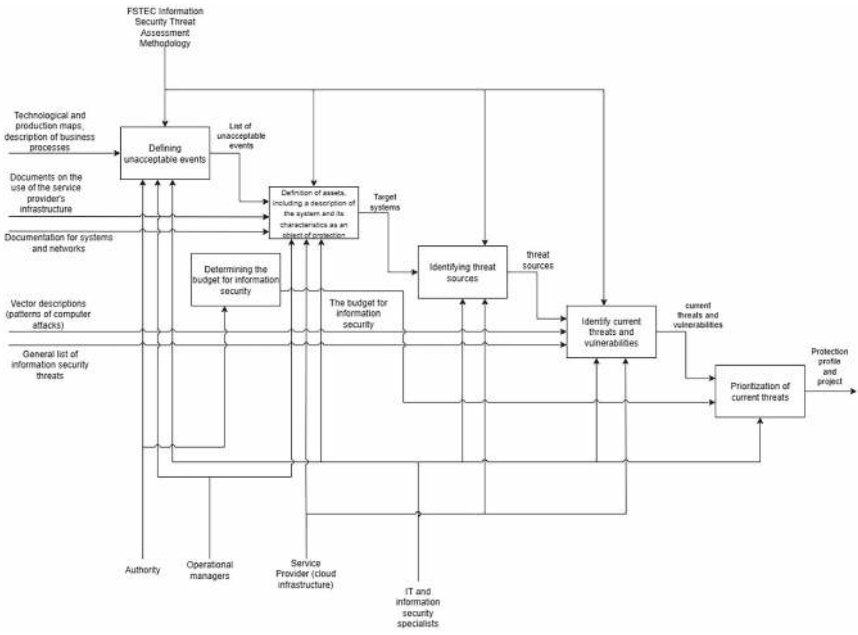


Fig. 1. A formalized representation of the information security risk assessment model using the graphical notation IDEF0.

Upon completion of the work within the proposed information security risk assessment model, a detailed information protection plan and project will be developed.

3.2 Improved Methods for Assessing Information Security Risk, Considering the Level of Vulnerability Risk

Let us denote the vulnerability risk assessment according to CVSS 3.0 as the actual value V, which ranges from 0 to 10. In order to maintain consistency with the traditional approach, we have normalized this parameter within the proposed framework to the range of 0 to 1 by dividing V by 10. This results in the following formula for calculating information security risk based on the cost of damage D_i and the risk of vulnerability V_i (1):

$$R = \sum_i D_i * \frac{V_i}{10}. \tag{1}$$

To account for the variability in risk over time, it is possible to consider the intensity of the accumulation of destructive effects on the system, if necessary $\lambda_i(t)$ (2):

$$R = \sum_i D_i * \frac{V_i}{10} * \lambda_i(t). \tag{2}$$

The methods described above may not always be suitable for accurately assessing information security risks, since this number does not take into account the relationship between the potential damage from a threat and the actual damage it can cause to an organization. Therefore, in some cases it may be more rational to consider risk as a combination of two linguistic fuzzy variables (2):

$$R = langle \mathrm{D}_i, V_i rangle, \tag{3}$$

where:

– D = "Damage", for which:
– universal set $U_D = [0, 10]$,
– term-set T(D) = {"missing", "low", "medium", "high", "critical"},
– semantic rule M for determining the correspondence of each fuzzy variable $\tilde{D}$ to the fuzzy subset M($\tilde{D}$) of the universal set U_D is presented in Table 3.

V = "CVSS", for which:

– universal set $U_V = [0, 10]$,
– term-set T(V) = {"missing", "low", "medium", "high", "critical"},
– semantic rule M for determining the correspondence of each fuzzy variable $\tilde{V}$ to the fuzzy subset M($\tilde{V}$) of the universal set U_D is presented in Table 4.

Table 3. The correspondence between the values of linguistic and fuzzy variables D.

Values of linguistic variables	Values of fuzzy variables
Missing	0.0
Low	0.0–4.0
Medium	4.0–7.0
High	7.0–9.0
Critical	9.0–10.0

However, this approach does not take into account important parameter: the cost of fix vulnerability. If the damage caused by the implementation of a vulnerability is less than the cost of fixing it, it is not always rational to fix this vulnerability. In order to take this metric into account, a new linguistic variable C = "Cost" is created with parameters similar to the variable D = "Damage", which is added to the vector R (4):

$$R = langle \mathrm{D}_i, V_i, C_i rangle \tag{4}$$

Table 4. The correspondence between the values of linguistic and fuzzy variables V.

Values of linguistic variables	Values of fuzzy variables
Missing	0.0–2.0
Low	2.0–4.0
Medium	4.0–6.0
High	6.0–8.0
Critical	8.0–10.0

The meaning of each linguistic variable V is as follows:

- "Missing"—there is no danger.
- "Low"—a violation of confidentiality, integrity, and accessibility may have limited adverse consequences for an organization or individual.
- "Medium"—a violation of confidentiality, integrity, and accessibility is likely to have a negative impact on an organization or individual.
- "High"—a violation of confidentiality, integrity, and accessibility is likely to have a significant negative impact on an organization or individual.
- "Critical"—a violation of confidentiality, integrity, and accessibility is likely to have a critical negative impact on an organization or individual.

4 Discussion

As part of the study, an experiment was conducted to test the individual stages of the proposed model for assessing information security risks. The experiment was based on a typical network architecture for an organization, which included the following assets:

- O.1—An automated workplace;
- O.2—Server;
- O.3—Peripheral equipment;
- O.6—Active network equipment;
- O.11—Information (data) stored in systems and networks;
- O.12—Physical communication lines.

The goal of the experiment was to determine which assets in a typical organization's network architecture have the highest potential for information security risk, based on various types of losses.

To do this, we used initial data for each asset in the organization, ranging from 1 to 3 risks, represented as a CVSS v3 vector (see Table 5). For each risk, a qualitative assessment was made of the potential cost of damage, ranging from "Low", "Medium", or "High". The results of experiment No. 2 are presented in Table 6.

The visualization of the experimental results is presented in the form of an accumulation diagram (see Fig. 2), in the form of a dot plot (see Fig. 3).

Table 5. Vulnerabilities of the organization's assets

Assets	Vector CVSS v3	Damage
O.1	AV:L/AC:H/PR:L/UI:N/S:U/C:N/I:N/A:H	Low
	AV:N/AC:H/PR:L/UI:R/S:C/C:L/I:N/A:H	High
O.2	AV:A/AC:H/PR:L/UI:N/S:C/C:H/I:H/A:H	Medium
O.3	AV:L/AC:H/PR:L/UI:R/S:C/C:L/I:N/A:H	Low
O.6	AV:N/AC:H/PR:L/UI:N/S:C/C:H/I:N/A:H	High
O.11	AV:N/AC:H/PR:N/UI:R/S:C/C:H/I:H/A:H	Low
	AV:L/AC:H/PR:H/UI:R/S:C/C:H/I:N/A:L	Low
	AV:P/AC:H/PR:L/UI:R/S:C/C:H/I:H/A:N	Medium
O.12	AV:L/AC:H/PR:H/UI:R/S:C/C:N/I:H/A:H	High

Table 6. Experimental results

Assets	Vector CVSS v3	Damage	CVSS assessment	Risk assess-ment
O.1	AV:L/AC:H/PR:L/UI:N/S:U/C:N/I:N/A:H	Low	4,8	2,76
	AV:N/AC:H/PR:L/UI:R/S:C/C:L/I:N/A:H	High	6,4	9,8
O.2	AV:A/AC:H/PR:L/UI:N/S:C/C:H/I:H/A:H	Medium	8,0	4,72
O.3	AV:L/AC:H/PR:L/UI:R/S:C/C:L/I:N/A:H	Low	6,1	1,96
O.6	AV:N/AC:H/PR:L/UI:N/S:C/C:H/I:N/A:H	High	8,2	6,9
O.11	AV:N/AC:H/PR:N/UI:R/S:C/C:H/I:H/A:H	Low	8,3	0,06
	AV:L/AC:H/PR:H/UI:R/S:C/C:H/I:N/A:L	Low	5,8	1,99
	AV:P/AC:H/PR:L/UI:R/S:C/C:H/I:H/A:N	Medium	6,5	3,52
O.12	AV:L/AC:H/PR:H/UI:R/S:C/C:N/I:H/A:H	High	6,9	6,53

Based on the data collected, the following conclusions can be drawn:

- Due to the high CVSS score of asset O.2 and its low cost of damage, the overall model score for this asset is quite low.
- Asset O.1 requires the most attention according to the assessment model, as it has the highest total assessment, with high CVSS indicators and a significant level of damage.
- Asset O.3 is the least important according to the model, with the lowest total estimation.

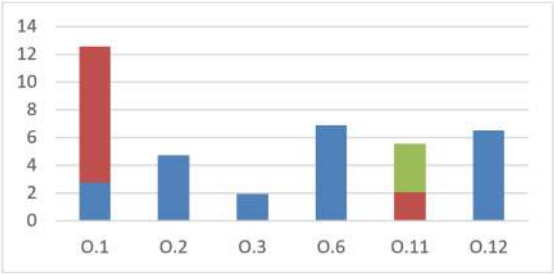


Fig. 2. The diagram of the total information security risk assessment on the assets of the organization

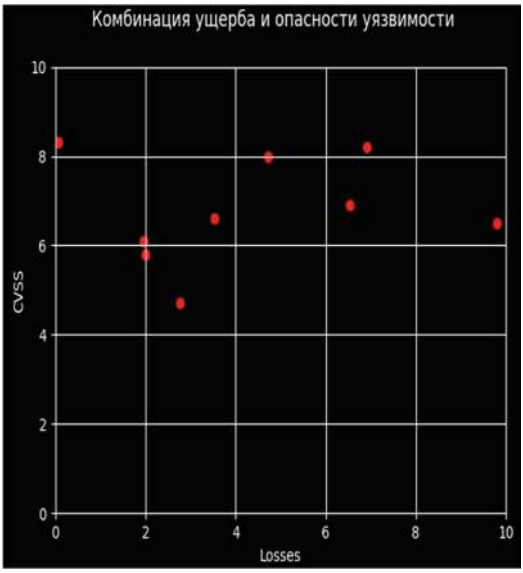


Fig. 3. A dot plot of the experiment results

5 Conclusion

From a scientific perspective, the potential of this research lies in further refining the developed model, which would allow for a deeper understanding of the subject matter. Additionally, the introduction of artificial intelligence (AI) [29] could improve the accuracy of results and simplify the process of working with the model, which has significant practical implications in terms of limited resources.

Practically speaking, the aim of this project is to create a decision support system that can assist in assessing the risk of information security threats.

References

1. National project Data Economy and Digital Transformation of the State. http://government.ru/rugovclassifier/923/about/. Last accessed 20 Feb 2025
2. Bagheri, S., Ridley, G., Williams, B.: Organisational cyber resilience: management perspectives. Australas. J. Inf. Syst. **27** (2023). https://doi.org/10.3127/ajis.v27i0.4183
3. Yanushkina, Y., Komkov, A.: Analysis of foreign experience in ensuring the stability of the national currency. Collection of scientific papers "digitalization of science and education: prospects for creating a unified knowledge system". In: Materials of the International Scientific and Practical Events of the Society of Science and Creativity, pp. 304–311. Kazan (2024)
4. Adel Alqudhaibi, Majed Albarrak, Sandeep Jagtap, Nikki Williams, Konstantinos Salonitis, Securing industry 4.0: assessing cybersecurity challenges and proposing strategies for manufacturing management, Cyber Secur. Appl. **3**, 100067 (2025). ISSN 2772–9184, doi:https://doi.org/10.1016/j.csa.2024.100067.
5. Komane, K., Khoza, L., Radebe, F.: A conceptual framework for cybersecurity awareness. J. Cyber Secur. **7**(1), 79–108 (2025). https://doi.org/10.32604/jcs.2025.059712
6. Carmi, E., Yates, S.: What do digital inclusion and data literacy mean today? Internet Policy Rev. **9**(2) (2020). https://doi.org/10.14763/2020.2.1474
7. Gu, Z., Xu, G., Hu, N.: Preface: security and safety in network simulation and evaluation. Secur. Saf. **4**, E2025004 (2025). https://doi.org/10.1051/sands/2025004
8. Taskin, N., Özkeleş Yıldırım, A., Ercan, H.D., Wynn, M., Metin, B.: Cyber insurance adoption and digitalisation in small and medium-sized enterprises. Information. **16**(1), 66 (2025). https://doi.org/10.3390/info16010066
9. Utenkova, M., Maksimova, E.: Forecasting the development of information security events in the context of information warfare. In: Lapina, M., Raza, Z., Tchernykh, A., Sajid, M., Zolotarev, V., Babenko, M. (eds.) AISMA-2024: International Workshop on Advanced Information Security Management and Applications. AISMA 2024 Lecture Notes in Networks and Systems, vol. 863. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72171-7_32
10. Maksimova, E., Utenkova, M.: Assessment of interobject interaction in the infrastructural system of the Armed Forces of the Russian Federation. The state and prospects of development of modern science in the field of "Information technologies in the Armed Forces of the Russian Federation". In: Collection of the III All-Russian Scientific and Technical Conference, pp. 33–36. Military innovation technopolis "ERA", Anapa (2024)
11. Tian, T., Zhang, C., Jiang, B., et al.: Insider threat detection for specific threat scenarios. Cybersecurity. **8**, 17 (2025). https://doi.org/10.1186/s42400-024-00321-w
12. Slapnicar, S., Vidmar, T., Tsen, E.: Process theory of supplier cyber risk assessment. Australas. J. Inf. Syst. **29** (2025). https://doi.org/10.3127/ajis.v29.5323
13. Zhukov, A., Tomin, N., Kurbatsky, V., Sidorov, D., Panasetsky, D., Foley, A.: Ensemble methods of classification for power systems security assessment. Appl. Comput. Inform. **15**(1), 45–53. ISSN 2210-8327, (2019). https://doi.org/10.1016/j.aci.2017.09.007
14. McNett, J., McNett, J.: Implementing a cybersecurity continuous user evaluation program. J. Cyber Secur. **7**(1), 279–306 (2025). https://doi.org/10.32604/jcs.2025.067514
15. Renney, H., Chenchiah, I.V., Nethercott, M., Paligadu, R., Lang, J.: An open and adaptable approach to vulnerability risk scoring. J. Cyber Secur. **7**(1), 221–238 (2025). https://doi.org/10.32604/jcs.2025.064958
16. Achaal, B., Adda, M., Berger, M., et al.: Study of smart grid cyber-security, examining architectures, communication networks, cyber-attacks, countermeasure techniques, and challenges. Cybersecurity. **7**, 10 (2024). https://doi.org/10.1186/s42400-023-00200-w
17. Fernandez, E.B., Yoshioka, N., Washizaki, H., et al.: Abstract security patterns and the design of secure systems. Cybersecurity. **5**, 7 (2022). https://doi.org/10.1186/s42400-022-00109-w

18. Muhammad, A.H., Hanafi, Yuana, K.A., Ghozali, B., Haris, R.: Data-driven security assessments for predicting information security maturity levels. J. Interner Serv. Inf. Secur. (JISIS). **15**(2 (May)), 906–925. https://doi.org/10.58346/JISIS.2025.I2.060
19. Ao, S., Fang, B., Xiao, X., Zhang, H.: Threat ripple model: a model to characterize business-oriented attacks based on business dependencies. Secur. Saf. **4**, 2025003 (2025). https://doi.org/10.1051/sands/2025003
20. Basan, A.S., Basan, E.S., Lapina, M.A., Kormakova, V.N., Lapin, V.G.: Security methods for a group of mobile robots according to the requirements of Russian and foreign legislation. IOP Conf. Ser.: Mater. Sci. Eng. **873**(1) (2020)
21. Parfenov, D., Bolodurina, I., Lapina, M.: Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Conf. Ser.: Mater. Sci. Eng. **873**, 012020 (2020). https://doi.org/10.1088/1757-899X/873/1/012020
22. Maksimova, E.A., Lapina, M.A., Lapin, V.G., Rusakov, A.M.: Anthropomorphic model of states of subjects of critical information infrastructure under destructive influences. In: Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I. (eds.) Mathematics and its Applications in New Computer Systems. MANCS 2021 Lecture Notes in Networks and Systems, vol. 424. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97020-8_51
23. Maksimova, E., Lapina, M., Lapin, V.: Synthesis of Models for Ensuring Information Security of Subjects of Critical Information Infrastructure under Destructive Influences. CEUR Work. Proc. **3094**, 108–117 (2022) https://ceur-ws.org/Vol-3094/paper_11.pdf
24. Common Vulnerability Scoring System. https://www.first.org/cvss/. Last accessed 20 Feb 2025
25. Ruohonen, J.: A look at the time delays in CVSS vulnerability scoring. Appl. Comput. Inform. **15**(2), 129–135. ISSN 2210-8327 (2019). https://doi.org/10.1016/j.aci.2017.12.002
26. Kainulainen, S., Tuunanen, T., Vartiainen, T.: Requirements risk Management for Continuous Development: Organisational needs. Australas. J. Inf. Syst. **28** (2024). https://doi.org/10.3127/ajis.v28.4441
27. Freier, P., Schumann, M.: Decision support Systems in the Context of cyber-physical systems: influencing factors and challenges for the adoption in production scheduling. Australas. J. Inf. Syst. **25** (2021). https://doi.org/10.3127/ajis.v25i0.2849
28. Utenkova, M., Komkov, A., Maksimova, E., DSS architecture for information security risk assessment. Current problems of applied mathematics, computer science and mechanics. In: Proceedings of the International Scientific Conference, Voronezh, December 2–4, Voronezh. C. 1402–1406 (2024).
29. Ferrag, M.A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., Debbah, M.: Generative AI in cybersecurity: a comprehensive review of LLM applications and vulnerabilities. Internet Things Cyber-Phys. Syst. **5**, 1–46 (2025). ISSN 2667-3452. 10.1016/j.iotcps.2025.01.001

Modified Transformer Architectures for Sports Action Recognition from Pose Sequences

Swetha V[1(✉)], Amos Bortiew[1], Aruna S K[1], and Stuart Barnes[2]

[1] Department of AI and Data Science Engineering, CHRIST University, Bengaluru, India
swetha.v@mtech.christuniversity.in,
{amos.bortiew,aruna.sk}@christuniversity.in

[2] Centre for Computational Engineering Sciences, Cranfield University, Cranfield, UK
s.barnes@cranfield.ac.uk

Abstract. In sports analytics, healthcare, surveillance, and various other fields, recognizing human action is essential. But issues like partial obstruction, varying viewpoints, and changes in light conditions frequently affect accurate object recognition. Most current approaches reduce robustness by treating pose estimation and action recognition as separate processes. We address this limitation by using our Modified PoseFormer and Modified MixSTE models to propose a unified pose-driven framework that combines both stages. After a comparative evaluation of pose estimation models, MoveNet Lightning was chosen for its speed and reliability in extracting 2D keypoints from the video. The proposed transformer models receive these keypoints, convert them into temporal pose sequences and then finally performed action recognition. Upon experimenting with the benchmark Penn Action dataset, it was found that our proposed approach surpassed the traditional state-of-the-art methods. This highlights the potential of our proposed work in enhancing real-world sports action recognition and video understanding based on poses.

Keywords: Human action recognition · Pose estimation · Movenet lightning · Movenet thunder · MediaPipe · PoseFormer · MixSTE

1 Introduction

A fundamental task in computer vision is recognizing human action in videos, since it has various applications in surveillance, sports analytics, and human-computer interaction. Identifying human activities accurately from videos enables intelligent systems to interpret and respond to real-world environments more effectively.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 708–717, 2027.
https://doi.org/10.1007/978-3-032-27834-0_62

Most traditional appearance-based methods rely on handcrafted motion features or raw pixels, making them sensitive to background noise, viewpoint differences, and lighting changes. Eventually, this degrades their performance in unconstrained environments where temporal and spatial modeling of human motion is required. Conversely, human pose estimation provides a structured representation of body movements, which is particularly useful for recognizing sports action. But most existing approaches handle pose estimation and action recognition separately, resulting in the propagation of errors. This emphasizes the need for a unified pipeline that integrates pose estimation with downstream action recognition, enhancing end-to-end learning and improved robustness.

To extract precise and accurate 2D keypoints from video frames, we evaluate pose estimation models in this work, including MoveNet Lightning, MediaPipe, and MoveNet Thunder. After performing a comparative analysis, we determined that MoveNet lightning was the most accurate and reliable model for keypoint detection. Then, we transformed these extracted keypoints into temporal pose sequences that serve as inputs to our proposed Modified PoseFormer and Modified MixSTE models for sports action recognition. In contrast to prior works that use transformer models only for pose lifting, we employ two proposed transformer models directly for action recognition by integrating into them a classification head. This feature enables models to capture spatiotemporal dependencies in skeletal data more effectively, achieving superior performance compared to existing state-of-the-art methods. This study provides the following significant contributions:

- A single pipeline that ensures reliable and effective performance by integrating pose estimation and action recognition.
- A comparative analysis of state-of-the-art pose estimation models to identify the most accurate and reliable pose estimator.
- Two transformer-based architectures, Modified PoseFormer and Modified MixSTE, are proposed for direct sports action recognition.

Experimental validation on the Penn Action dataset demonstrates that the proposed Modified PoseFormer and Modified MixSTE outperform the current leading models.

2 Related Work

Human pose estimation and action recognition techniques are the two primary areas typically employed in human action recognition research. It has been understood that human pose estimation has been recognized as a primary approach for studying and modeling human movement. Various models, including MoveNet Lightning, MoveNet Thunder, and MediaPipe Pose, have been widely used for extracting 2D keypoints from RGB video frames. MoveNet Lightning provided a real-time, efficient solution with a good balance between speed and accuracy. On the other hand, MoveNet Thunder prioritized accuracy and robustness at the

expense of high computational cost. Furthermore, MediaPipe Pose gave integrated tracking to identify consistent and trustworthy keypoints [1–4]. In spite of these advancement, there remained challenges on applying these techniques due to complex environments found in the data.

In addition to lightweight methods, recent research works have found that 3D pose estimation and sequence modeling using transformer-based architectures are moving beyond lightweight 2D pose estimation. For 3D pose estimation, Zheng et al. [5] introduced a spatial-temporal transformer that captures both the inter-frame temporal dependencies and intra-frame relations between joints. While MixSTE [7] introduced a mixed spatio-temporal encoder to boost sequence modeling, PoseFormer [6] built on this concept to describe long-range dependencies in pose sequences.PoseFormerV2 [8] applied frequency domain representation for long skeletal sequences to increase efficiency and robustness. Although transformer-based models had been primarily used for pose estimation rather than action recognition, these studies show the potential of these models.

Significant progress has also been observed in human action recognition, particularly with techniques that use skeletal data. In order to extract spatial-temporal features from pose sequences for action recognition, Xu et al. [9] introduced ActionFormer, a model that combines CNN and transformer modules. With a focus on datasets, problems, and deep learning methods, Wang and Yan [10] and Zhang et al. [11] provided a detailed analysis of RGB and skeleton-based action recognition. In order to identify human actions, Sun et al. [12] first proposed a two-stage pose estimation model to extract skeleton keypoints, then performed feature modeling and classification. While Fu et al. [13] optimized YOLOv7-Pose for fitness movement recognition, Feng and Meunier [14] gave a description of skeleton graph neural network-based methods for action recognition. On regular datasets, vision transformer-based models have also been applied directly for action recognition [15], with reasonable accuracy.

Several earlier approaches for action recognition, such as Ji [17], who investigated and studied a pose-based action recognition framework; Kulkarni et al. [18], who extracted discriminative features from video frames using MobileNet and transfer learning; and DeepPear [16], which combines skeletal poses with appearance features around predicted joints. Temporal dependencies from skeletal sequences have been captured by applying sequence modeling techniques, such as 1D-CNNs [19]. Lastly, [20] showed how important both spatial and temporal features are for reliable performance.

The studies from the literature show that pose estimation and action recognition are often treated as separate processes. Even though transformer-based models, like PoseFormer and MixSTE, have shown outstanding results in pose estimation, there is still very little application of them in direct action recognition. Furthermore, the long-range spatio-temporal dependencies in pose sequences are not utilized fully by basic action recognition techniques, which either use raw RGB frames or pre-extracted skeletal features. In regard to these limitations, our work provides a unified solution framework that brings together action recognition and pose estimation. To give a single pipeline that combines

accurate pose estimation with improved action recognition ability, a comparative analysis of the best models among the state-of-the-art pose estimators is performed in our approach, and then the keypoints extracted from best pose estimator is transformed into pose sequences. Then, direct action recognition is performed by adapting transformer-based pose estimation models into Modified PoseFormer and Modified MixSTE architectures to enhance action recognition performance.

3 Methodology

3.1 Framework Overview

A unified framework is proposed for human action recognition, which utilizes extracted keypoints from video frames. The pipeline mainly integrates three key components: pose estimation, temporal sequence construction, and transformer-based action recognition. Figure 1 below represents a unified pipeline proposed for recognizing human actions based on pose data.

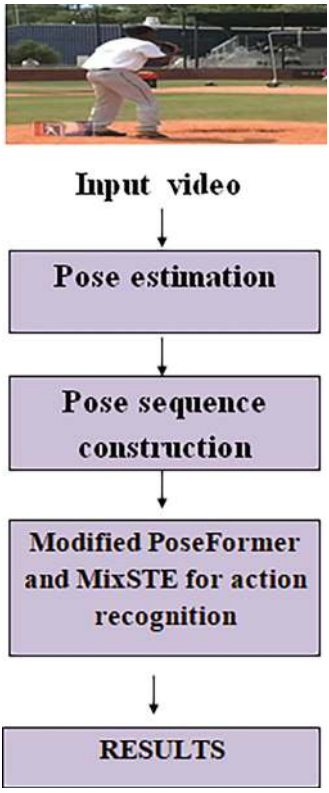


Fig. 1. Proposed framework for human action recognition using pose sequences

Initially, from each video frame, the 2D keypoints are extracted from state-of-the-art pose estimation models, such as MoveNet Lightning, MoveNet Thunder, and MediaPipe. Based on a comparative evaluation of these models, the best pose estimator is selected, and the extracted keypoints from this model are transformed into temporal pose sequences, as the dynamics of human motion are encoded in temporal pose sequences. The modified transformer-based models proposed, Modified Poseformer and Modified MixSTE, receive these pose sequences as input, which then categorize actions into the appropriate groups after learning spatiotemporal dependencies. This unified pipeline offers a reliable end-to-end solution for pose-based action recognition by guaranteeing that temporal dynamics and spatial configurations are efficiently captured.

3.2 Pose Estimation

Human pose estimation serves as the foundational step for action recognition. The goal is to locate keypoints corresponding to human joints in each video frame. For a given frame f_t, the pose estimator outputs J 2D keypoints as follows:

$$P_t = \{(x_{t,1}, y_{t,1}), (x_{t,2}, y_{t,2}), \ldots, (x_{t,J}, y_{t,J})\}, \tag{1}$$

where P_t represents the pose at frame t, J denotes the number of joints, and $(x_{t,j}, y_{t,j})$ are the 2D coordinates of the j-th joint. Collecting P_t across all frames generates a temporal sequence of poses that captures both spatial joint configuration and motion information.

We evaluated three state-of-the-art 2D pose estimation models: MoveNet Lightning, MoveNet Thunder, and MediaPipe Pose to select the best pose estimator for our proposed unified pipeline. Among these, MoveNet Lightning provided the most efficient and reliable keypoints, which serve as robust input for temporal sequence construction.

3.3 Temporal Pose Sequence Construction

MoveNet Lightning was chosen as the best pose estimator after a comparative analysis of the state-of-the-art pose estimation models. To understand the dynamics of human motion over time, the key points extracted from this model are then transformed into temporal pose sequences. The main idea behind the construction of temporal pose sequences is to organize the extracted key points across successive frames to create continuous motion trajectories. The temporal sequence of a video with T frames can be written as follows:

$$S = [P_1, P_2, \ldots, P_T], \tag{2}$$

where $S \in \mathbb{R}^{T \times J \times 2}$ encodes the 2D coordinates of every joint over time. This representation is more resilient to background clutter and different points of view because it preserves spatial configurations and records temporal dependencies.

3.4 Transformer Based Action Recognition

After constructing temporal pose sequences, the next step is to classify human actions. To capture spatio-temporal dependencies, we propose two transformer-based architectures: Modified PoseFormer and Modified MixSTE. These models were initially designed for pose estimation and are adapted here for direct action recognition by adding a classification head.

Modified PoseFormer The temporal sequence S is embedded with positional information as shown in Eq. (3), which is then processed through L transformer layers in Eq. (4):

$$Z^0 = S + E_{\text{pos}}, \tag{3}$$

$$Z^l = \text{PoseFormerLayer}(Z^{l-1}), \quad l = 1, 2, \ldots, L, \tag{4}$$

where E_{pos} represents learnable positional embeddings. The final embedding Z^L is then mapped to action probabilities:

$$\hat{y}_{\text{PF}} = \text{Softmax}(\text{Classifier}(Z^L)). \tag{5}$$

Algorithm 1 Modified PoseFormer for Action Recognition

Require: Pose sequence S, number of layers L
Ensure: Predicted action label $\hat{y}_{\text{PF}}$
1: $Z^0 \leftarrow S + \text{PositionalEmbedding}$
2: **for** $l = 1$ **to** L **do**
3: $\quad Z^l \leftarrow \text{PoseFormerLayer}(Z^{l-1})$
4: **end for**
5: $\hat{y}_{\text{PF}} \leftarrow \text{Softmax}(\text{Classifier}(Z^L))$
6: **return** $\hat{y}_{\text{PF}}$

Modified MixSTE Similarly, the input sequence is embedded and processed through L alternating spatial-temporal layers and then mapped to action probabilities as follows:

$$Z^0 = S + E_{\text{pos}}, \tag{6}$$

$$Z^l = \text{MixSTELayer}(Z^{l-1}), \quad l = 1, 2, \ldots, L, \tag{7}$$

$$\hat{y}_{\text{MX}} = \text{Softmax}(\text{Classifier}(Z^L)). \tag{8}$$

Algorithm 2 Modified MixSTE for Action Recognition

Require: Pose sequence S, number of layers L
Ensure: Predicted action label $\hat{y}_{\text{MX}}$
$Z^0 \leftarrow S + \text{PositionalEmbedding}$
for $l = 1$ **to** L **do**
  $Z^l \leftarrow \text{MixSTELayer}(Z^{l-1})$
end for
$\hat{y}_{\text{MX}} \leftarrow \text{Softmax}(\text{Classifier}(Z^L))$
return $\hat{y}_{\text{MX}}$

Algorithms 1 and 2 summarize the proposed approach, showing how Modified PoseFormer and Modified MixSTE process temporal pose sequences with added classification heads to predict human actions. They highlight the key contribution: adapting transformer-based pose estimation models for direct, end-to-end action recognition while jointly leveraging spatial and temporal attention.

4 Experiments and Results

4.1 Dataset and Preprocessing

The Penn Action dataset, which consists of annotated video clips covering 15 sports activities, was used for the experiments. To extract 2D skeletal keypoints, each video was handled frame-by-frame. The temporal and spatial patterns of human motion were then captured by transforming these keypoints into fixed-length temporal sequences. To prepare for scale and position variations, the sequences were normalized. To ensure that every action class was represented in both sets, a stratified train-test split was applied.

4.2 Pose Estimation: Comparative Analysis

For pose estimation, three models were evaluated: MoveNet Lightning, MoveNet Thunder, and MediaPipe Pose. Evaluation metrics included the number of keypoints detected per frame, model size, total inference time across five representative videos, and average frames per second (FPS). Table 1 represents the comparative analysis of state-of-the-art pose estimation models.

Table 1. Comparative analysis of three pose estimation models

Model	Keypoints/ Frame	Model size (MB)	Total inference Time (s)	Average FPS
MoveNet lightning	17	7	16.54	20.3
MoveNet thunder	17	14–15	73.96	5.3
MediaPipe pose	33	90	53.96	7.3

Table 1 shows that MoveNet Lightning offers the most favorable trade-off between speed and efficiency among the evaluated models. It is also considerably lightweight (∼7 MB) and integrates easily with the action recognition pipeline. Therefore, based on these metrics, MoveNet Lightning was chosen as the optimal pose estimator for subsequent action recognition tasks.

4.3 Action Recognition Performance

Two proposed modified transformer-based models, Modified PoseFormer and Modified MixSTE, were used by the proposed action recognition framework to classify actions directly from 2D pose sequences. The evaluation results of the proposed modified transformer models are compared with the state-of-the-art action recognition methods as shown in Table 2.

Table 2. Action recognition performance

Model	Accuracy (%)	Precision	Recall	F1-Score
1D-CNN	76.50	0.78	0.77	0.77
MobileNet	73.33	0.74	0.73	0.73
PoseFormer	83.26	0.85	0.83	0.83
MixSTE	81.12	0.82	0.81	0.81

From the table, it is observed that the proposed transformer-based models outperform traditional action recognition methods such as MobileNet and 1D-CNN in terms of accuracy and F1-score, as shown in Table 2. The Modified PoseFormer with the highest F1-score shows a strong ability to capture long-range spatiotemporal dependencies in pose sequences. By efficiently encoding mixed spatiotemporal features, the Modified MixSTE enhances the ability to recognize intricate motion patterns.

4.4 Qualitative Results

Along with quantitative metrics, a qualitative evaluation provides deeper insights into how well the proposed models capture fine-grained human motion dynamics. Example frames from the Penn Action dataset with 2D pose keypoints overlaid, along with the corresponding ground truth and predicted action labels, are illustrated in Fig. 2. The qualitative analysis demonstrates the models' strong ability to generalize across diverse sports actions, complementing the numerical performance metrics.

(a) PoseFormer (b) MixSTE

Fig. 2. Qualitative results of the proposed modified transformer models for action recognition

5 Conclusion

One of the major tasks in computer vision is recognizing human actions in videos. In this study, we present a unified framework for temporal pose sequence-based sports action recognition. MoveNet Lightning was found to be the best model for effective keypoint extraction after a comparison of the state-of-the-art pose estimation techniques. In order to effectively model human actions, temporal pose sequences were developed from the best pose estimator model's extracted keypoints to function as structured motion descriptors. These pose sequences were fed into the proposed transformer-based architectures, Modified PoseFormer and Modified MixSTE, which gave better performance than traditional techniques, showing that they could be used to provide a more robust pose-based action recognition system. Overall, the outcomes demonstrate how well the proposed unified pipeline combines pose estimation and action recognition and provides an enhanced pose-based action recognition system within a comprehensive end-to-end framework.

References

1. Jo, B., Kim, S.: Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices. J. Theor. Appl. Inf. Technol. **39**, 1–11 (2023). https://doi.org/10.18280/ts.39011
2. Salisu, S., Mohamed, A.S.A., Jaafar, M.H., Pauzi, A.S.B., Younis, H.A.: A survey on deep learning-based 2D human pose estimation models. CMC-Comput. Mater. Continua (2023). https://doi.org/10.32604/cmc.2023.035904
3. Parashar, D., Mishra, O., Sharma, K., Kukker, A.: Improved yoga pose detection using MediaPipe and MoveNet in a deep learning model. Revue d'Intelligence Artificielle **37**(5), 1197–1202 (2023). https://doi.org/10.18280/ria.370511
4. Chung, J.-L., Ong, L.-Y., Leow, M.-C.: Comparative analysis of skeleton-based human pose estimation. Future Internet **14**, 380 (2022). https://doi.org/10.3390/fi14120380
5. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3D human pose estimation with spatial and temporal transformers. In: Proceedings of IEEE/CVF ICCV, pp. 11656–11665 (2021). https://doi.org/10.1109/ICCV48922.2021.01146

6. Zheng, Y., Zhu, Y., Huang, J., Yang, Y.: PoseFormer: Direct human pose estimation with transformers. In: Proceedings of IEEE/CVF ICCV, pp. 746–755 (2021). https://doi.org/10.1109/ICCV48922.2021.00077
7. Zhang, C., Li, X., Xu, W., Liu, Y., Lu, C.: MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video. In: Proceedings of IEEE/CVF CVPR, pp. 13232–13242 (2022). https://doi.org/10.1109/CVPR52688.2022.01304
8. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation. In: Proceedings of IEEE/CVF CVPR, pp. 8877–8886 (2023). https://doi.org/10.1109/CVPR52729.2023.00857
9. Xu, C., Zhou, Y., Zhu, A., Wang, Z., Li, Y., Hu, F.: ActionFormer: pose-based action recognition via transformer and CNN. In: 2022 China Automation Congress (CAC), pp. 36–41 (2022). https://doi.org/10.1109/cac57257.2022.10055176
10. Wang, C., Yan, J.: A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. IEEE Access, 11, 53880–53898 (2023). https://doi.org/10.1109/access.2023.3282311
11. Zhang, Y., You, S., Karaoglu, S., Gevers, T.: 3D human pose estimation and action recognition using fisheye cameras: a survey and benchmark. Pattern Recognit. **141**, 111334 (2024). https://doi.org/10.1016/j.patcog.2024.111334
12. Sun, R., Zhang, Q., Guo, J., Chai, H., Li, Y.: Human action recognition using skeleton data from two-stage pose estimation model. In: 2021, pp. 769–779. https://doi.org/10.1007/978-3-030-89095-7_73
13. Fu, H., Gao, J., Liu, H.: Human pose estimation and action recognition for fitness movements. Comput. Graph. **116**, 418–426 (2023). https://doi.org/10.1016/j.cag.2023.09.008
14. Feng, M., Meunier, J.: Skeleton graph-neural-network-based human action recognition: a survey. Sensors **22** (2022). https://doi.org/10.3390/s22062091
15. Fahim, M., Arefin, M.: A vision transformer based model for human action recognition. In: 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–5 (2025). https://doi.org/10.1109/ecce64574.2025.11013387
16. Jhuang, Y.Y., Tsai, W.J.: Deeppear: Pose-based deep representations for action recognition. In: Proceedings of 25th ICPR, pp. 7119–7126 (2020). https://doi.org/10.1109/ICPR48806.2021.9413011
17. Ji, J.: Action recognition based on human pose estimation. In: 2021 2nd International Conference on Computing and Data Science (CDS), pp. 73–78 (2021). https://doi.org/10.1109/CDS52072.2021.00019
18. Kulkarni, S.S., Jadhav, S.: Insight on human activity recognition using deep learning approach. In: 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 1–5 (2023). https://doi.org/10.1109/ESCI56872.2023.10099759
19. Srimath, S., Ye, Y., Sarker, K., Sunderraman, R., Ji, S.: Human activity recognition from RGB video streams using 1D-CNNs. In: 2021 IEEE SmartWorld, Ubiquitous Intelligence and Computing, pp. 295–305 (2021). https://doi.org/10.1109/SWC50871.2021.00048
20. Wani, M.H., Faridi, A.R.: Deep learning-based video action recognition: a review. In: 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 243–249 (2022). https://doi.org/10.1109/ICCCIS56430.2022.10037736

Automated System for Monitoring a Hero's Image on the VKontakte Social Network Using Artificial Intelligence

Ilya Zavyalov, Milada Krapivina(✉), Oleg Gabrielyan, Mikhail Gasparyan, and Ivan Kravchenko

V.I. Venadsky Crimean Federal University, Simferopol, Russia
miladakrapivina@gmail.com

Abstract. The article presents the development and implementation of an automated hero image monitoring system on the VKontakte social network using artificial intelligence methods. The proposed service combines an artificial intelligence module, a server-side web application, and a client-side web interface, which provides ease of use, architectural flexibility, and scalability. The DeBERTa model was used as the basis for text analysis, which showed the best results among the considered analogues. The system collects, processes and classifies data on a daily basis, providing researchers with the opportunity to build statistical graphs, track the dynamics of discussions and conduct a qualitative analysis of texts. The results obtained demonstrate the effectiveness of using deep learning methods to study socio-cultural phenomena in the digital environment.

Keywords: Client-server architecture · Web service · Machine learning · Text classification · Hero image · Web-scraping · Social networks · DeBERTa.First section

1 Introduction

In today's information society, social networks are becoming a crucial platform for shaping public opinion and cultural perceptions [1]. Among them, the narrative of the hero occupies a special place—a symbolic image that reflects the values and guidelines of society. Tracking how and in what contexts users discuss heroes (whether historical figures, participants in military events, representatives of working professions, or characters from personal stories) is of considerable interest for research in the fields of sociology, political science, and cultural studies.

Social research is particularly relevant now [2], when the turnover of posts on social networks is extremely high. At the same time, manual processing of the huge amounts of text information generated in them is extremely laborious and inefficient. This leads to a scientific and practical deficit: there are no automated methods for analyzing and classifying such complex and multifaceted phenomena as the image of a hero. The Russian language is difficult to process using traditional text analysis tools, and sociologists

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 718–727, 2027.
https://doi.org/10.1007/978-3-032-27834-0_63

and linguistic researchers need to monitor not only official statistics, but also collect and interpret data themselves in accordance with historical events.

We offer a solution that will allow sociologists and philologists to conduct research on linguistic constructions and their evolution using the image of the hero as an example. Data is collected from the social network VKontakte, as this platform is one of the most important in Russia and the CIS countries, with the largest number of users. The analysis of VKontakte should provide a fairly multifaceted view of the phenomena occurring in society and language [3].

During the development of the software complex, a modular architecture was designed, including three key components: an artificial intelligence subsystem, a server web application, and a user web interface. This structural division ensures the independent development of individual modules and also contributes to the scalability and flexibility of the system in conditions of growing computational loads and expanding functionality. The system was created taking into account the best practices at the time of writing, with detailed consideration of the architecture to achieve the best future expandability of the project, as well as to simplify the task of deploying the web application to the server.

2 Methods

2.1 Neural Network

To ensure high-quality text classification, a model based on DeBERTa [4, 5] was trained. In addition to this model, LSTM (RNN) and LightAutoML (AutoML) were also considered. The model was selected from among its counterparts as the most accurate, with the best f1-macro score. Figure 1 shows the classification report.

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.51	0.65	77
1	0.00	0.00	0.00	0
2	0.06	0.33	0.10	6
3	0.77	0.52	0.62	46
4	0.07	0.25	0.11	4
accuracy			0.50	133
macro avg	0.36	0.32	0.30	133
weighted avg	0.80	0.50	0.60	133

Fig. 1. Classification report for DeBERTa.

Figure 2 shows the Normalized Confusion Matrix.

A dataset of 977 posts was collected from the Russian social network VKontakte in 2022, using keyword searches for variations of the word "hero" (e.g., "герой", "герои").

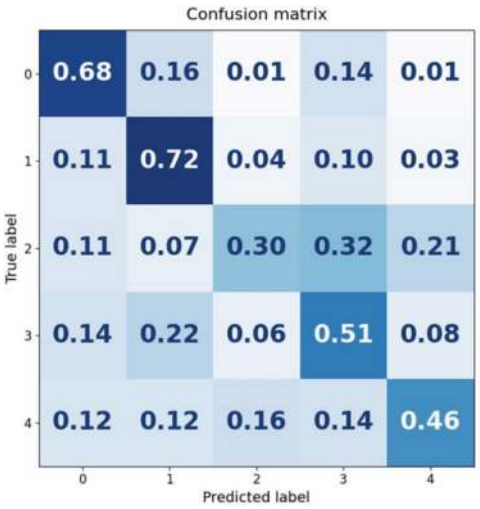


Fig. 2. Normalized confusion matrix for DeBERTa.

After removing duplicates (primarily reposts), the final dataset comprised 828 unique posts. Only the textual content of each post was used for classification, with metadata (e.g., date, author) excluded. The resulting dataset has been made publicly available for research purposes [6].

The matrix confirms improved performance, with diagonal values (correct predictions) dominating. However, misclassifications persist between Classes 2 (labor/MCHS/police) and 4 (personal context), with ~48 instances of Class 4 predicted as Class 2 and ~40 instances of Class 2 as Class 4. This reflects lexical overlap (e.g., "rescue," "heroic act" in both). Military classes (0, 1, 3) showed clearer separation, particularly Class 1 (WWII), with only ~15–20 misclassifications out of ~251 instances, due to distinct markers (e.g., "veteran," "1941–1945"). Classes 0 (SVO) and 3 (other conflicts) had minor errors (~58 instances of Class 0 as Class 3, ~40 vice versa), attributed to shared military terminology.

The refined class structure and balanced dataset reduced confusion between military contexts and improved recognition of personal narratives (Class 4, recall = 0.67). However, distinguishing professional roles (Class 2) remains challenging, suggesting the need for additional features (e.g., named entity recognition) or larger datasets.

2.2 Architecture of the System

The server-side web application is implemented in the client-server paradigm, which is a common approach in the creation of modern distributed web systems [7]. The client part is developed based on the Single Page Application (SPA) concept using the React library and TypeScript language. This choice of technologies minimizes the number of requests to the server, increases the interactivity of the user interface, and ensures smooth interaction with the system [8].

The server part is built using the ASP.Net Core platform and implements a RESTful API that provides standardized access to the system's functionality. The main tasks of this module are processing requests from the client application, validating and routing data, and controlling user access rights, i.e., ensuring security. The PostgreSQL database management system is used to store information, and interaction with it is carried out using Entity Framework Core, which provides convenient integration of object-oriented models with a relational data structure.

In addition to supporting the main functionality of the web application, the server module performs tasks for extracting data from external sources. In particular, integration with the VK social network has been implemented through the official API, which allows searching for news messages by specified keywords and time intervals. The data obtained undergoes additional processing using a neural network, which performs the task of content classification. The classification results are stored in a database and serve as the basis for subsequent statistical analysis, visualization, and identification of dynamic patterns in information flows.

The assembled front-end application is presented as a set of minified static files (HTML, CSS, JavaScript) served by the NGINX web server. The latter also acts as a reverse proxy and processes requests to the server application. To ensure data transfer security, HTTPS protocol support is implemented using certificates that are automatically generated and updated by the Let's Encrypt system.

When designing the system, special attention was paid to ensuring scalability, ease of maintenance, and automation of deployment processes. All core services (database, web server, server application, and neural network module) are packaged in isolated Docker containers, which simplifies dependency management and allows the execution environment to be reproduced on different computing platforms. Containers are managed using Docker Compose, which provides declarative description and quick initialization of the entire system.

To automate the build, test, and deployment processes, a GitLab-based continuous integration and delivery (CI/CD) system has been implemented. This approach minimizes the likelihood of human error, speeds up the release of updates, and contributes to more stable and predictable operation of the software complex.

The neural network is located in a separate Python service using FastAPI. The service receives requests from the backend and processes the received data using DeBERTa. The resulting classification is sent back as an HTTP response. This service is closed to external access outside the server.

Full architecture is provided on Fig. 3

A certificate of intellectual property registration (No. 2025685562) was obtained for the server implementation and architecture.

3 Results

As a result of this work, a service was created and deployed on the server of V.I. Vernadsky KFU. It is currently operational and already helping sociologists and philologists to analyze data. Figure 4 shows the appearance of the system's main page after authorization:

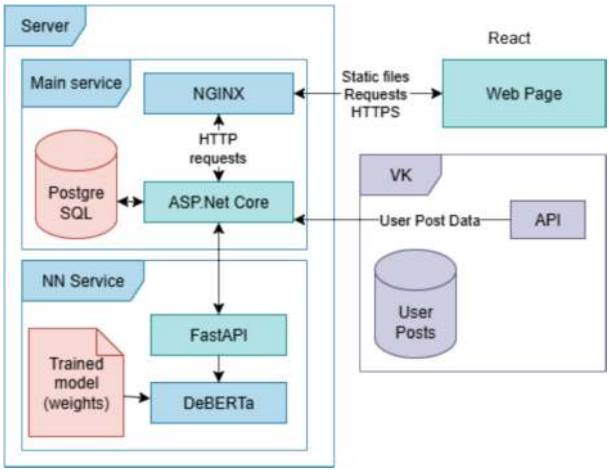


Fig. 3. Server software architecture.

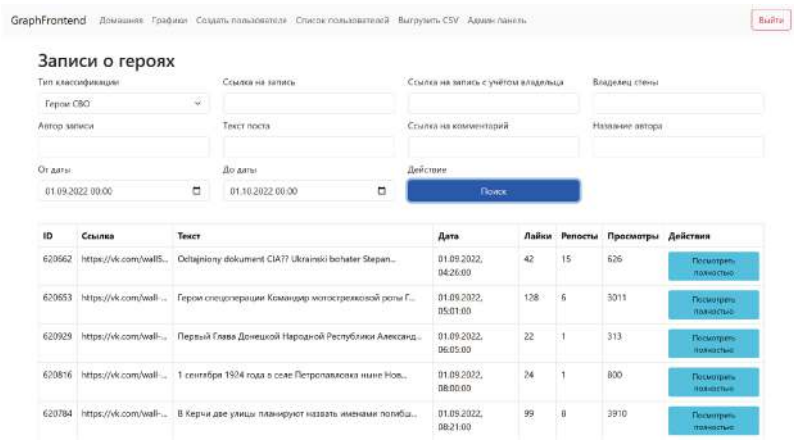


Fig. 4. Home page of the service—a database access point.

The service allows you to search by classification type, author name, and the text of the post itself. Another important feature is the ability to limit your search by date (from and to). You can view the posts you find and get detailed information about each one, including the number of views, likes, and reposts, as well as the group and person who posted it.

The data can be used to build graphs that aid in analysis. Figure 5 shows an example of a bar graph that shows the exact number of posts in a given classification for a given period:

In addition to a bar chart, you can construct a more visual pie chart, as shown in Fig. 6.

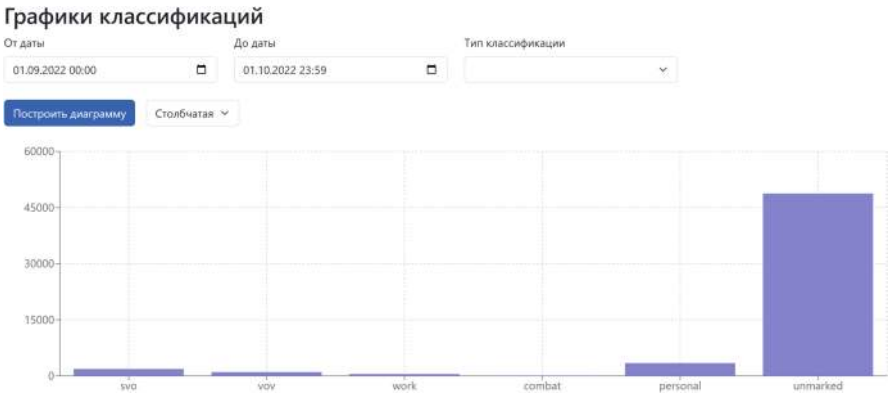


Fig. 5. Post bar graph.

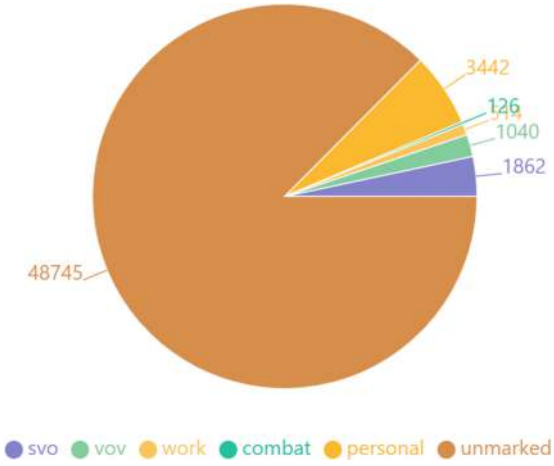


Fig. 6. Example pie chart.

In addition to charts, the system can generate graphs based on view rate (VR), engagement rate (ER), and average rating (Fig. 7).

The system collects information about new posts from the social network VKontakte every day and adds them to the database.

Fig. 7. VR, ER and average rating graphs.

4 Discussions

The results obtained have already been actively used by researchers, confirming the demand for the developed service in the scientific community. In addition, the system architecture is designed with scalability and transferability to the computing power of other research teams in mind. This ensures flexibility in distribution and increases the prospects for its further use. At the same time, during development, both the positive aspects of the implemented solutions and certain limitations were identified that require attention in subsequent iterations of the system's development.

First, the use of modern technological approaches made it possible to significantly speed up the development process and quickly make the changes needed by researchers. C# and ASP.NET Core made it possible to create a flexible infrastructure for the backend server. To enforce flexibility of the codebase a Clean Architecture (also known as "Onion Architecture") implementation was used. Since this system was developed with various research needs in mind it was quite important to ensure that the software would scale well in the future.

Second, the artificial intelligence module was the core component to increase and speed-up the quality of text analysis. Unlike traditional text processing methods, a neural network is able to deeply grasp the semantic context of the sentences and, as a result, unwrap complex meanings hidden in the text. This allowed to confirm how truly useful neural networks are when it comes to analyzing complex messages, such as social media posts.

Despite these achievements, the system has a number of limitations that require further refinement:

First, the current model is focused exclusively on analyzing content related to a specific research context (e.g., a particular historical or sociological figure). Attempting to use the service to analyze other topics will require modifying the module for loading data from VK, as well as retraining the model. Additionally, the implementation of ensemble approaches, including the use of multiple neural networks or a combination with classical machine learning algorithms, is considered promising [9]. This approach will increase the system's resilience to different types of input data and expand its analysis capabilities.

Second, data analysis is limited to posts on the VKontakte social network, which provides an incomplete picture of the information field. In particular, comments that may contain important details for identifying additional contexts and clarifying semantic relationships are not taken into account. In addition, alternative data sources, such as Odnoklassniki or Yandex Zen, which could significantly expand the representativeness of the sample, are not integrated. In the future, it would be advisable to consider multi-channel data collection to form a more comprehensive view of the phenomena under study.

Third, the system has limited practical value for the general user. The specificity of the collected data and the results of the analysis are primarily of interest to the research community (historians, sociologists, political scientists, and media communications specialists). This circumstance significantly complicates the commercialization of the product and emphasizes its academic and scientific research applications.

Fourth, the model is limited to the Russian language, which reduces its versatility. Posts on VKontakte often contain publications in English, Belarusian, Ukrainian, and other languages, which reduces the quality of the analysis [4]. A promising direction for further research is to expand the linguistic capabilities of the model, including the introduction of multilingual architectures and the use of pre-trained language models (multilingual transformers), which will provide broader coverage of input data and increase the accuracy of analysis.

Further development of the service involves several key areas that will overcome the identified limitations and expand its scope of application:

One of the priority areas is to expand the language capabilities of the system. The transition from a monolingual model to a multilingual architecture based on modern transformer models (mBERT, XLM-R, GPT-like multilingual models) will allow analyzing publications not only in Russian, but also in English, Ukrainian, Belarusian, and other languages [10]. This is especially relevant in the context of the globalization of the information space and the need to take into account intercultural differences in communication.

Expanding the range of analyzed information sources is a promising area. In addition to VKontakte, integration with other social networks and media platforms such as Odnoklassniki, Yandex Zen, as well as international services (e.g., Twitter/X or Telegram). Multi-channel data collection will allow for a more representative picture and increase the accuracy of identifying social and cultural trends.

Contemporary research in the field of social network analysis goes beyond purely textual data. In this regard, there are plans to develop the system towards multimodal analysis, including the processing of images, videos, and user comments. This approach

will not only clarify the context of messages, but also reveal hidden patterns related to the visual aspects of communication [10].

Ensemble methods (i.e. combining several neural network architectures and/or classical machine learning algorithms) are to be considered for an improvement in analysis accuracy. This was the model will be much more resistant to noise data, reducing the number of classification errors.

The service has value among the researchers in sociology, humanities experiments, as well as media discourse analysis. It could be considered as a tool to aid in these fields. Automating data collection and automatic analysis will allow researches to not only acquire old data, but to collect it daily and investigate new facts, while removing the technical costs of doing so manually.

Lastly, it would be beneficial to continue research towards making this service faster and lighter by using distributed computing and cloud platforms. Not only will this enable the system to process much larger volumes of data, it will be able to provide access to a wide range of research groups, regardless of their technical abilities and infrastructure.

5 Conclusion

The service was able to demonstrate high efficiency when it comes to analyzing and classifying text from the VK social network. The codebase, created with scalability and modularity in mind, allowed the system to be ready for various research contexts and teams. The integration of deep learning methods was able to greatly improve accuracy when compared to a traditional method as well as save time when compared to manual processing.

But the system is not without its limits. In its current state, the service is primarily focused on analyzing posts only, ignoring the comments that could provide additional valuable context. It is also fair to say that the reach of the system is currently quite limited, as it's only polling data from VK, ignoring other social networks. The neural model works with Russian language only, which reduces the universality of the analysis. To overcome these limitations, directions for further development have been proposed: expanding language support, integrating additional data sources, introducing multimodal analysis (including images and videos), and applying ensemble methods to improve the stability and accuracy of classification.

Thus, the presented service is a significant tool for sociologists, philologists, and media discourse researchers, opening up new opportunities for analyzing the dynamics of the hero's image in the modern media space.

Acknowledgement. The research was carried out at the expense of a grant from the Russian National Science Foundation (project No. 24-28-00413 "The Image of a hero in the modern Russian media space: state and prospects").

References

1. Ausat, A.M.A.: The role of social media in shaping public opinion and its influence on economic decisions. Technol. Soc. Perspect. TACIT **1**(1), 35–44 (2023). https://doi.org/10.61100/tacit.v1i1.37

2. Social Media Popularity: A Currency for Social Power, in ResearchGate. https://www.researchgate.net/publication/369113764_Social_Media_Popularity_A_Currency_for_Social_Power. Accessed 29 Sept 2025
3. VK Company: Consolidated Annual Report 2023 (ENG). https://corp.vkcdn.ru/media/files/VK_AR2023_ENG_06.06.pdfVKCorp (2024)
4. CMC: DeBERTa-GRU: Sentiment Analysis for Large Language Model. https://www.techscience.com/cmc/v79n3/57133. Accessed 29 Sept 2025
5. Paper page—A Family of Pretrained Transformer Language Models for Russian. https://huggingface.co/papers/2309.10931. Accessed 29 Sept 2025
6. Russian hero context dataset: Kaggle. https://www.kaggle.com/datasets/miladakrap/russian-hero-context-dataset. Accessed 29 Sept 2025
7. Nyabuto, G.: Client-server architecture, a review. Int. J. Adv. Sci. Comput. Appl. **3**(2) (2024). https://doi.org/10.47679/ijasca.v3i1.48
8. Modern Front End Web Architectures with React.Js and Next.Js: Semantic Scholar. https://www.semanticscholar.org/paper/Modern-Front-End-Web-Architectures-with-React.Js-Fariz-Lazuardy/9df33541508d8ee667593bb630532f4fc9a52af6 (2025). Accessed 29 Sept 2025
9. Ensemble multifeatured deep learning models and applications: a survey. IEEE J.S Mag. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10265042. Accessed 29 Sept 2025
10. Automated visual analysis for the study of social media effects: opportunities, approaches, and challenges. Commun. Methods Meas. **18**(2).https://doi.org/10.1080/19312458.2023.2277956. Accessed 29 Sept 2025

Hybrid Dequantized Machine Learning Models for Information Security Incident Detection

Sergei Zuev(✉) and Victor Trifonov

V.I. Vernadsky Crimean Federal University, Simferopol, Russia
zuevsv@cfuv.ru

Abstract. This paper considers the issue of recognizing cybersecurity incidents in an event stream. The factors determining the complexity of the task are indicated. It is proposed to use modern solutions in the field of data analysis, obtained from the field of quantum-inspired machine learning. The basic architecture of an intelligent incident recognition system using a dequantized classifier and an artificial neural network as a parameter selector for the classifier is presented. The indicators of the classifier's precision and the efficiency of information security measures are compared. It is shown that the use of dequantized classifiers as part of an intelligent incident recognition system can make preventive information protection measures effective.

Keywords: SIEM · Dequantized classifier · Quantum-inspired machine learning · Dequantized model

1 Introduction

1.1 Difficulties with Building an Intellectual SIEM System

The complexity of intelligent analysis of information security events is due to at least three reasons. Firstly, the number of incidents in the event stream is extremely small compared to normal events–the two classes are disproportionately represented in the data. Secondly, in the topological space of event features, it is very difficult to determine a distance correlating with incident patterns–there is no metric in the space. Thirdly, an intelligent incident detection system must adapt to changes in the environment, and do this in real time–thus, online training is required. In principle, modern architectures of intelligent systems have the means to overcome all three obstacles [1]. At the same time, there are not only computational, but also business-motivated requirements for any potential solution that tackles the problem of recognizing suspicious events. The application of the solution should be economically justified and not lead to a large number of false positives, and it also should not be excessively expensive to develop, implement, or operate.

One result from the field of research of quantum machine learning methods are so-called dequantized models [2]. They are constructed in a feature space which, in structure, coincides with the space of quantum states of some quantum system. But regular classical

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 728–734, 2027.
https://doi.org/10.1007/978-3-032-27834-0_64

data structures can be derived from the quantum description and then be processed normally on classical hardware. It is known that dequantized machine learning models differ favorably from classical ones in that they learn faster and allow incremental, or online, learning [3]. These are precisely the qualities that can help overcome difficulties outlined above with the least expenditure of resources and, quite possibly, with the best result.

1.2 Background and Related Works

There are already a number of works attempting to integrate intelligent analyzers into the workings of SIEM and IDS systems. Among the proposed solutions are feed-forward neural networks [4–7], recurrent neural networks [7, 8], convolutional neural networks [9], algorithms utilizing logical regression [10], naive Bayesian classifiers [11], and other approaches, such as certain hybrid systems [12, 13]. Most of these papers demonstrate a high Accuracy metric of 80% or more.

Due to the fact that information security incidents account for a minimal fraction of the total number of events, the standard Accuracy metric almost always displays high values, even if the classifier is performing poorly. This is deceptive. For example, if there are only 500 incidents in a collection of 100 000 records, Accuracy will easily reach 99% even with 1 000 false positives. The ensuing security measures taken can be very costly to the information system.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

With this in mind, the Precision metric would be more useful, since it takes into account the number of false positives:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

For this reason, it is necessary to evaluate the results of these works in more detail, rather than just focusing on the high Accuracy metric.

Another problem is the dependence of many machine learning architectures on labeled data. When using architectures, such as RNNs or optimized BPNNs, frequent retraining is required to adapt to new attack vectors. In addition, any retraining requires expensive and time-consuming human expertise. Although some approaches attempt to alleviate this by using a teacher model. Given that the nature of information security threats is changing very quickly, such data requirements are far from ideal. Even research aimed at analyzing and optimizing the cost functions and the learning process of controlled neural networks [4, 6] cannot completely eliminate the fundamental dependence on high-quality and relevant data labels.

There are relatively fewer works exploring architectures that allow for unsupervised learning. Several studies exist which utilize autoencoder models and achieve Precision over 80% [14–16]. But these works are based on detecting attacks rather than all incidents. They are also based on the NSL-KDD dataset [17], the share of attacks in the training sample of which reaches 46.5%. Such a high proportion of attacks does not correspond to the actual distribution of events in production SIEM or IDS systems.

Experiments have shown that if the training sample contains a small number of incidents–about 0.5%, and such rarity is indeed characteristic of the real data considered in this paper, then the best current neural network-based classifiers achieve a Precision score no higher than 5–6%. Therefore, their use is ineffective.

It should be noted that using the Precision metric also has its drawbacks, since the number of undetected incidents is not taken into account, while being a critical characteristic of an SIEM or IDS system. Skipping a real attack due to low Recall can have disastrous consequences. The ideal option for assessing the balance between false positives and undetected incidents is the F-Score metric, which represents the harmonic mean of Recall and Precision.

$$F_1 = 2\frac{Recall \cdot Precision}{Recall + Precision} = \frac{2TP}{FN + FP + 2TP} \tag{3}$$

However, this metric, while most effectively reflecting the practical applicability of a model in a security environment, is rarely published in literature.

1.3 Objective of This Work

The purpose of this work is to develop a basic architecture of a dequantized classifier for recognizing information security incidents and methods for integrating it into other intelligent systems, with the possibility of utilizing a hybrid machine learning model, to ensure satisfaction of enterprise interests.

2 Dequantized Classifiers

2.1 The Principle Behind a Dequantized Classifier

A model in which data is represented as states of a certain quantum system is called a quantum model. If instead of a physical system its mathematical description is used, and a quantum computer is not required for data processing, then such a model can be called dequantized [2]. It can be constructed by mapping the feature space onto a Hilbert space of the desired dimension, that is, the number of dimensions of the feature space of the input data and the Hilbert space of states of the quantum system must be equal. From a quantum computing point of view, the dimension of the Hilbert space has the form B^n, where B is the number of basic states in the simplest system ($B = 2$ for qubits, and $B > 2$ for qudits), and n is the number of simplest systems in a composite system [18]. It is clear that the number of features in real data will not always match the quantum system. Therefore, it is often necessary to encode data into a different number of dimensions while not losing any information. One of the methods of such dimensionality reduction is implemented in the genser software module (available on PyPI) and allows the user to change the dimension of an integer data set in a wide range [3],p.179.

The representation of data in the form of a quantum state mathematically looks as such:

$$(x_0, \ldots, x_{N-1}) \rightarrow |q\rangle = \sum_{k=0}^{N-1} a_k |k\rangle, \tag{4}$$

where $N = 2^n$ (in the qubit case), $M = 2^{n+1} - 2$ and

$$a_0 = cosx_0 \ldots cosx_{\frac{M}{2}-1} \tag{5}$$

$$a_k = sinx_{k-1} cosx_k \ldots cosx_{\frac{M}{2}-1} e^{ix_{\frac{M}{2}+k-1}}, k = 1, \ldots \frac{M}{2} - 1, \tag{6}$$

$$a_{\frac{M}{2}} = sinx_{\frac{M}{2}-1} e^{ix_{M-1}}. \tag{7}$$

This is the data encoding process for the dequantized model. The model itself is a combination of Bell states dependent on the specific task. In the case of a recurrent architecture, the model may exhibit algorithmic interconnection [18], p.26.

The training of the dequantized model is done by collecting measurement statistics on the training data. These measurements will produce statistical distributions that are suitably different for different classes in the input data. Thus, the question of whether a given events belongs to a given class can be answered by comparing the typical statistical distribution of that class to the distribution of measurement results of the event.

Since, by its construction, the learning process of the dequantized model does not involve backpropagation, it naturally implements online learning.

2.2 A SIEM Architecture Utilizing a Dequantized Classifier

The use of these models in real intelligent systems is conditioned by the possibility of their integration into existing infrastructure. Currently, information security events are processed using SIEM technology, which can monitor, analyze, and manage security events. It may contain various analyzers, which can be either signature-based or heuristic. Intelligent analyzers have also been gaining popularity in recent years–among Russian examples, TIAS hardware and software complex from the Infotecs company could be cited [19]. The embedded classifier should ensure the effective application of security measures.

Let's say, for example, that the acceptable risk for a certain critical resource is R_a, and the usual risk when the resource is running is R_0. The corresponding incident probabilities are then derived like so: $P_a = \frac{R_a}{C}$, $P_0 = \frac{R_0}{C}$. If we denote the probability that the predicted incident is indeed an incident as p, then protective measures should be taken when $p > P_a$. If the cost of the measures is M and they reduce the damage from the incident by a factor of $\mu < 1$, then the condition for a given p becomes:

$$\mu Cp > CP_a - M \tag{8}$$

($M < CP_a$ here, since measures cannot cost more than the damage caused by the realization of an acceptable risk. If this is the case, then criticality was assessed incorrectly), otherwise the security measures are ineffective. It follows that p is always larger than $P_a = \frac{R_a}{C}$. In realistic estimates, the value p (Precision in binary classification) should exceed 15–20% [20].

Augmenting the SIEM architecture with a dequantized classifier can significantly increase Precision, and when using online learning, it is possible to constantly adapt

the classifier to the environment, controlling the quality of its training. To do this, one can use a dequantized classifier as the main deciding mechanism, while using a neural network to optimize its hyperparameters and other characteristics, including the width of the learning window, the dimensionality of inputs and outputs, and the specifics of logical inference for comparing statistical distributions (see Fig. 1.).

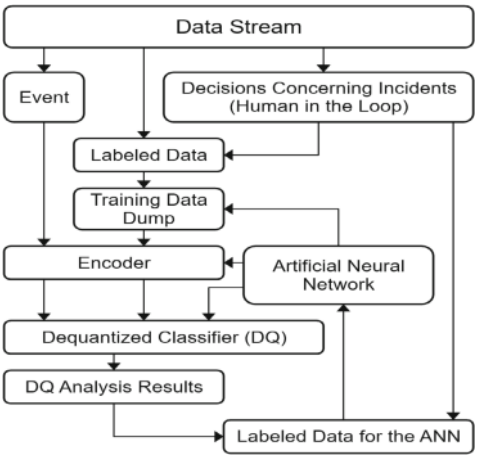


Fig. 1. A SIEM architecture utilizing a dequantized classifier managed by an ANN.

When using such an architecture, the increased Precision metric–which in itself may still be insufficient for the effective application of measures according to the criteria already described–can be improved to a satisfactory level by adjusting the parameters of the window width, the dimensionality of the working state space of the quantum system, as well as the method of logical inference on distributions. Preliminary software models have shown the promise of the architecture proposed: the Precision value increases significantly with the correct parameters.

3 Conclusion

Modern developments in the field of quantum-inspired machine learning methods, particularly dequantized classifiers, can be used in detecting information security incidents in an event stream. It is possible to combine a dequantized classifier with an artificial neural network in order to obtain an incident recognition accuracy value above the necessary threshold required for the effective application of information security measures. This paper presents a possible architecture of such an intelligent system.

Acknowledgements. The authors would like to express their gratitude to Kliver LLC and Infosecurity LLC for providing the opportunity to conduct computational experiments.

References

1. Mokalled, H., Catelli, R., Casola, V., Debertol, D., Meda, E., Zunino, R.: The applicability of a siem solution: requirements and evaluation. In: IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 132–137 (2019)
2. Tang, E.: Dequantizing algorithms to understand quantum advantage in machine learning. Nature Reviews Physics. **4**(11), 692–693 (2022)
3. Zuev, S.: Statistical online learning in recurrent and feedforward quantum neural networks. Dokl. Math. **108**(2), 317–324 (2023)
4. Ingre, B., Yadav, A.: Performance analysis of NSL-KDD dataset using ANN. In: Proceedings of the International Conference on Signal Processing and Communication Engineering Systems (SPACES), Vijayawada, India (2013)
5. Chiba, Z., Abghour, N., Moussaid, K., El Omri, A., Rida, M.: A novel architecture combined with optimal parameters for back propagation neural networks applied to anomaly net- work intrusion detection. Computers and Security. **75**, 36–58 (2018)
6. El Hajji, S., Moukafih, N., Orhanou, G.: Analysis of neural network training and cost functions impact on the accuracy of IDS and SIEM systems. In: Codes Cryptology and Information Security, pp. 433–451. Springer International Publishing, Berlin, Germany (2019)
7. Ussath, M., Jaeger, D., Cheng, F., Meinel, C.: Identifying suspicious user behavior with neural networks. In: Proceedings of the 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), vol. 2017, New York, NY, USA
8. Yin, C., Zhu, Y., Fei, J., He, X.: A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access. **5**, 21954–21961 (2017)
9. Chouhan, N., Khan, A., Khan, H.U.R.: Network anomaly detection using channel boosted and residual learning based deep convolutional neural network. Appl. Soft Comput. **83**, 105612 (2019)
10. Gurung, S., Ghose, M.K., Subedi, A.: Deep learning approach on network intrusion detection system using NSL-KDD dataset. International Journal of Computer Network and Information Security. **11**(3), 8–14 (2019)
11. Ch, R., Gadekallu, T.R., Abidi, M.H., Al-Ahmari, A.: Computational system to classify cyber crime offenses using machine learning. Sustainability. **12**(10), 4087 (2020)
12. Bhattacharya, S., R, K.S., Maddikunta, P.K.R., et al.: A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU. Electronics. **9**(2), 219 (2020)
13. Swarna Priya, R.M., Maddikunta, P.K.R., Parimala, M., et al.: An effective feature engineering for DNN using hybrid PCA- GWO for intrusion detection in IoMT architecture. Comput. Commun. **160**, 139–149 (2020)
14. Javaid, A., Niyaz, Q., Sun, W., Alam, M.: A deep learning approach for network intrusion detection system. In: Proceedings of the 9th EAI International Conference on Bio-Inspired Information and Communications Technologies (Formerly BIONETICS), pp. 21–26 (2016)
15. Su, T., Sun, H., Zhu, J., Wang, S., Li, Y.: BAT: deep learning methods on network intrusion detection using NSL-KDD dataset. In: IEEe Access. **8**, 29575–29585 (2020)
16. Choi, H., Kim, M., Lee, G., Kim, W.: Unsupervised learning approach for network intrusion detection system using autoencoders. J. Supercomput. **75**(9), 5597 (2019)
17. Zhao, R.: NSL-KDD. IEEE Dataport. https://doi.org/10.21227/8rpg-qt98 (2022).
18. Zuev, S.: Geometric properties of quantum entanglement and machine learning. Russian Technol. J. **11**(5), 19–33 (2023)

19. Software and hardware complex for automatic detection of incidents based on the analysis of information security events ViPNet TIAS. https://infotecs.ru/products/vipnet-tias, Last accessed 05 May 2025
20. Ho, C.Y., Lai, Y.C., Chen, I.W., Wang, F.Y., Tai, W.H.: Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems. IEEE Commun. Mag. **50**(3), 146–154 (2012)

Review of Cybersecurity-Aware Machine Learning System for Real-Time Fatigue Detection Based on Physiological and Behavioural Data

Munis Xamidov[1](✉), Otabek Esanov[1], and Manish Singla[2]

[1] Samarkand State University, Samarkand, Uzbekistan
samsungmunis@gmail.com
[2] Chitkara University, Rajpura, Punjab, India

Abstract. The growing global problem of road accidents due to fatigue-induced driver impairment creates a growing need for innovative and reliable systems for fatigue detection. Real-time fatigue detection methods using machine learning (ML) that utilize behavioral and physiological indicators such as eye closure, yawning, head position, and facial landmarks offer a promising non-invasive solution. However, the use of AI technologies in mobility remains a challenge, especially with new cybersecurity issues to protect sensitive biometric data collected and used in real time. This thesis describes an analysis of Cybersecurity-Aware Machine Learning (CA-ML) system for real-time driver fatigue detection that integrates facial landmark analysis with temporal behavioural patterns. We compared the performance of many ML architectures, including Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and CNN + Long Short-Term Memory (CNN + LSTM) networks, to achieve high classification accuracy with low latency. Furthermore, we investigated potential system vulnerabilities and security frameworks that include secure data gathering, liveness detection, and adversarial robustness to assure functional safety and data integrity. Results in analysed systems achieved 94.2% accuracy and a response time of 120 ms, proving its ability to improve driver safety while addressing important cybersecurity concerns in automotive AI applications. In this extended review we provide reproducible methodological details and quantitative evaluation guidelines for CA-ML systems. We present closed-form feature definitions, a recommended dataset split and preprocessing pipeline, and a comparative results table that contrasts SVM, pure CNN, and hybrid CNN + LSTM models under identical conditions. We also include a cybersecurity evaluation checklist—covering liveness detection, encrypted telemetry, and adversarial robustness testing—that can be immediately applied by practitioners and researchers.

Keywords: Driver fatigue detection · Cybersecurity-aware machine learning · CNN-LSTM architecture · Facial landmark analysis · Real-time monitoring · Adversarial robustness · Intelligent transportation systems

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 735–747, 2027.
https://doi.org/10.1007/978-3-032-27834-0_65

1 Introduction

1.1 Background

Driver fatigue is a serious public safety concern because it significantly impairs cognitive and motor functions, compromising driving performance. According to the World Health Organization (WHO), fatigue causes approximately 20% of all road traffic accidents, making it one of the leading causes of preventable deaths on global roads. Fatigue reduces reaction time, situational awareness, judgment, and the likelihood of microsleep episodes, which can be fatal at high speeds. According to the National Highway Traffic Safety Administration (NHTSA), drowsy driving causes more than 100,000 police-reported crashes in the United States each year, resulting in thousands of injuries and hundreds of deaths [1]. To combat this growing threat, modern vehicles are increasingly outfitted with AI-powered driver monitoring systems (DMS). These systems use in-cabin cameras, biometric sensors, and machine learning algorithms to detect early signs of fatigue and warn drivers before an accident occurs. However, despite advances in AI and sensor technologies, current fatigue detection systems have significant limitations in terms of detection accuracy and system security, particularly when deployed in real-world scenarios [2]. Chaos in the field of information security opens the way to multimillion-dollar losses, undermining customer trust and lawsuits [3–6]. Therefore, today, more than ever, it is important to give primary attention to information security.

1.2 Problem Statement

Modern fatigue detection systems rely heavily on visual cues such as eye closure (PERCLOS—Percentage of Eye Closure Over the Pupil over Time), head nodding, yawning frequency, and vehicle telemetry such as steering instability and lane departure. While these methods have shown moderate success in controlled environments, they suffer from several critical limitations:

- Visual-based systems have limited accuracy in real-world conditions due to poor lighting, occlusions (e.g., sunglasses, facial hair), and variations in driver appearance and posture.
- Lack of Early Detection: Fatigue symptoms often appear after a driver has reached a compromised state. Pure visual models do not capture subtle internal indicators like changes in heart rate variability or brainwave activity.
- Cybersecurity Vulnerabilities: Because these systems collect and process sensitive biometric data in real time, they are vulnerable to a variety of cyber threats such as spoofing attacks (for example, using photos or videos to circumvent facial recognition), adversarial manipulation of input data, and unauthorized access to stored or transmitted health information.

These limitations reduce the effectiveness of existing systems and raise concerns about both functional reliability and user privacy, especially as vehicles become more connected and integrated with cloud-based services.

1.3 Proposed Solutions

This thesis made a review of these issues and analysed a proposed Cybersecurity-Aware Machine Learning (CA-ML) system for real-time fatigue detection that combines multi-modal behavioural and physiological indicators with strong cybersecurity safeguards [7]. Unlike traditional systems that rely solely on visual inputs, proposed framework uses facial landmark tracking, temporal behavioural patterns, and, optionally, non-invasive physiological signals (such as heart rate variability and galvanic skin response) to improve detection accuracy. The system uses advanced deep learning architectures, such as Convolutional Neural Networks (CNNs) for spatial feature extraction from facial images and Long Short-Term Memory (LSTM) networks for temporal modelling of behavioural sequences, to detect both instantaneous and evolving signs of fatigue. Furthermore, it is incorporated that Support Vector Machines (SVMs) is good enough for comparative analysis and model robustness evaluation [8]. Crucially, the proposed system incorporates a comprehensive cybersecurity layer designed to protect against spoofing, adversarial attacks, and data breaches. This includes:

- Secure data acquisition protocols with hardware authentication,
- Biometric liveness detection to prevent presentation attacks,
- End-to-end encryption of biometric and behavioural data streams,
- Adversarial defense mechanisms to ensure model integrity.

By combining accurate fatigue classification with strong cybersecurity protections, the CA-ML system enhances both driver safety and data privacy, addressing two of the most pressing concerns in intelligent transportation systems today [9]. This thesis aims to review the real-time fatigue detection framework that integrates robust ML models and comprehensive cybersecurity strategies. The proposed system in this review thesis processes visual indicators—eye state, yawning frequency, and head posture—using facial landmark-based features while ensuring secure handling of sensitive biometric data throughout the pipeline [10].

1.4 Research Objectives and Contributions

Main purpose of this work is to create a design and do some analysis concerning cybersecurity and machine learning aspects of the system that can detect a fatigue in real time. While traditional methods mostly focus on systems that can function separate in terms of security and ML modelling, this suggested architecture integrates both. Following contributions specified in this work:

1. Single CA-ML Framework: The behavioural and physiological parameters are fused in a consolidated, steady, and privateness-preserved device the usage of CNN and LSTM structure.
2. Formalized Feature Computation: Key fatigue indicators together with Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and Head Posture Assessment have been mathematically described and standardized for computation.
3. Multi-Model Comparison: Assessment of the performance of SVM in contrast to CNN and CNN + LSTM inside the identical preprocessing and facts parameters [11].

4. Integrated Cybersecurity Assessment: Propose a modular cybersecurity checklist that during pairs considers: encryption, counterfactual robustness.
5. Pre-processing Pipelines and Dataset Partitioning: Design a device to partition and homogenize datasets for reproducibility to make sure a truthful foundation for assessment.

These contributions help to create a DDDS (driver drowsiness detection system) with high accuracy, secure architecture and transparency.

2 Related Work

Some related researchers have done this kind of system based on EEG data and considering rule-based approaches like PERCLOS, but still, there are some limitations and considerable issues. Mostly, systems that focus on data collection with EEG and EOG sensors, which are wearable tools, do not consider their inconvenience and sensitivity to real driving conditions. However, while PERCLOS-based methods are non-invasive, they can be computationally expensive and sensitive to environmental conditions such as lighting variations and occlusions [12]. To avoid this kind of limitations, this research focuses on a non-invasive, camera-based fatigue detection system that uses facial landmark analysis and temporal behavioural modelling. This method eliminates the need for intrusive hardware while maintaining high accuracy via intelligent feature extraction and deep learning. In addition, the cybersecurity-aware design principles are used to address emerging threats associated with AI-powered vehicle systems [13].

2.1 Limitation in Existing Systems

- Cybersecurity Vulnerabilities: Many current driver monitoring systems do not provide adequate protection against adversarial attacks, in which subtle changes in input data can mislead the model [14].
- Spoofing Attacks: Systems that rely solely on facial features are vulnerable to presentation attacks, which use photos, videos, or 3D masks to simulate alertness.
- Lack of Temporal Modelling: Most existing approaches treat each frame separately, failing to capture the progressive nature of fatigue, which develops over time.

The limitations described above can be seen with more cited works in the table below (Table 1):

Table 1. Citing existing papers in terms of limitations [15].

Authors	Year	Methodology	Accuracy (%)	Limitations
Zhang et al.	2018	Eye closure (PERCLOS)	87	Lighting dependency
Patel et al.	2020	EEG-based SVM classifier	91	Invasive sensors
Lee and Kim	2021	Steering behaviour + CNN	89	Lacks physiological integration

2.2 Proposed Solution

In this review, it can be seen that authors of reviewed research addressed these issues by combining deep learning models with temporal sequence modelling to better understand fatigue development. They specifically test and compare the performance of Convolutional Neural Networks for spatial feature extraction, Support Vector Machines for baseline classification, and CNN + LSTM architectures for capturing long-term dependencies in behavioural patterns over time [16]. The security-first framework with biometric liveness detection, adversarial training, and robust feature extraction to prevent spoofing and evasion attacks were also implemented.

Driver data collected from sensors is protocolled to ensure data transmission considers the privacy and integrity of this data. The combination of highly accurate drowsiness classification and security awareness in this system can solve most problems in monitoring the driver's state in smart vehicles.

3 System Architecture

3.1 System Components

This proposal envisions a real-time fatigue detection system using a camera. In capturing fatigue detection, deep learning will be employed for spatial and temporal modelling, along with security-centric processing. The system will truly be accurate and robust. The design includes four main components: Image Acquisition, Preprocessing, Feature Extraction, and Machine Learning Classification, as shown in Fig. 1 [17].

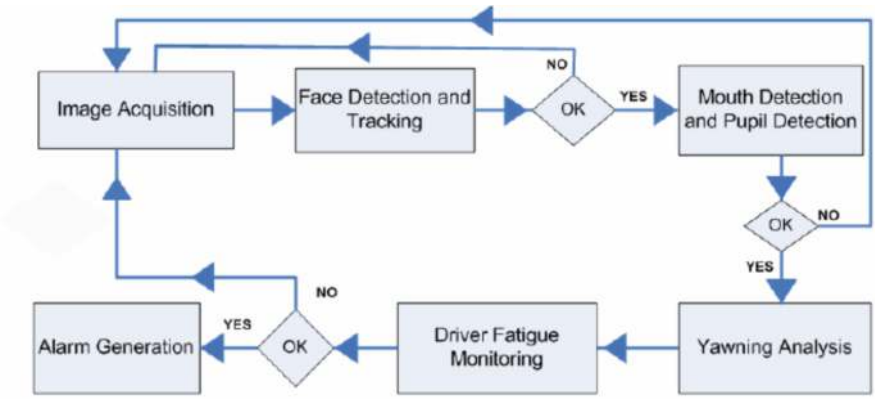


Fig. 1. The system architecture.

Key fatigue indicators derived from these landmarks include (Table 2):

Table 2. Fatigue indicators based on their computation method [18].

Fatigue indicator	Description	Computation method
Eye aspect ratio (EAR)	Measures eye openness	$EAR = \frac{\|P2-P6\|+\|P3-P5\|}{2\|P1-P4\|}$ (1)
Mouth aperture ratio (MAR)	Quantifies yawning	$MAR = \frac{\|m_3-m_9\|+\|m_4-m_8\|+\|m_5-m_7\|}{2\cdot\|m_2-m_6\|}$ (2)
Head pose estimation	Detects nodding behaviour	Using PnP (Perspective-n-Point) algorithm with 3D facial landmarks [7]

Head Pose (yaw, pitch, roll). Estimate using solvePnP with a standard 3D face model; the angular deviation magnitude θ can be summarized as: These features are extracted for each frame and aggregated into a time-series buffer to model behavioural dynamics over short intervals (e.g., 5–10 s).

$$\theta = \sqrt{\varphi_{\text{yaw}}^2 + \varphi_{\text{pitch}}^2 + \varphi_{\text{yaw}}^2} \tag{3}$$

Use moving-average smoothing (window = 1 s) to remove frame noise.

Temporal Aggregation. Let $f_t = [EAR_t, MAR_t, \theta_t]$ denote the feature vector at tim *t.* A temporal sequence $F = \{f_1, f_2, \ldots, f_T\}$ is fed to the LSTM for temporal modelling:

$$h_t = \text{LSTM}(f_t, h_{t-1}) \tag{4}$$

The final hidden state h_T represents the aggregated fatigue dynamics over the observation window.

Classification Decision. The fused CNN–LSTM model outputs the probability of fatigue as:

$$\text{P}(\textit{fatigue}) = \sigma(W_r[h_T \oplus c] + b) \tag{5}$$

where c represents CNN-extracted spatial features, ⊕ is concatenation, and σ is the sigmoid activation function (Table 3).

Table 3. Summarizes feature sensitivity based on experimental observations.

Feature	Physiological indicator	Sensitivity to fatigue	Temporal stability
EAR	Eye closure	High	Medium
MAR	Yawning	Medium	High
θ (Head pose)	Nodding	Medium	High

3.2 Preprocessing

- Face detection: use a lightweight detector (e.g. RetinaFace or MTCNN), then localize 68-point facial features (dlib or a modern alternative).
- Alignment: affine alignment to canonical features (eyes and nose) before cropping the patches.
- Temporal buffering: store a circular buffer of the last 5–10 s (150–300 frames at 30 fps) for temporal models.
- Normalization: normalize the image intensity in each channel (zero mean, unity variance) and increase the data: horizontal rotation, small rotation ($\pm 10°$), brightness/contrast jitter.

3.3 Machine Learning Classification

To classify the driver's state as either alert or fatigued, a hybrid CNN + LSTM deep learning architecture was employed that captures both spatial patterns and temporal evolution of fatigue symptoms [19].

– CNN Branch: A lightweight Convolutional Neural Network (CNN) such as MobileNetV2 or Tiny-YOLO is used to extract high-level spatial features from raw facial image patches or landmark heatmaps [20].
– LSTM Branch: The sequence of extracted EAR, MAR, and head posture values over time is fed into a Long Short-Term Memory (LSTM) network to model the progression of fatigue states [21].
– Fusion Layer: Features from both CNN and LSTM branches are concatenated and passed through a fully connected layer followed by a sigmoid activation function for binary classification (alert versus fatigued) [22].

Additionally, for comparison and interpretability, the evaluation of standalone models was done:

– SVM Classifier: Trained on engineered features (EAR, MAR, head angle) using RBF kernel.
– Pure CNN: For spatial-only classification.
– Pure LSTM: For temporal-only modelling.

This modular and scalable architecture enables real-time fatigue monitoring while maintaining resilience against spoofing and adversarial attacks through integration with cybersecurity mechanisms described in later sects [23].

4 Results

We evaluated three detector families under identical preprocessing and dataset splits. The test set contains 10,000 frames extracted from in-cabin recordings, with class balance 3,000 fatigued frames (30%) and 7,000 alert frames (70%). The compared models are (Table 4): an engineered-feature SVM (features: EAR, MAR, head-angle), a spatial-only CNN, and a fused CNN + LSTM (spatial CNN features concatenated with LSTM temporal encoding). Reported metrics are Accuracy, Precision, Recall, F1-score, AUC, and inference latency (mean $\pm$ std) measured end-to-end (face detection $\rightarrow$ landmark extraction $\rightarrow$ model forward-pass) on an RTX-5080-equivalent device (Table 4).

Table 4. Summary of model performance metrics on the test set: accuracy, precision, recall, F1-score, AUC, and measured inference latency.

Model	Accuracy(%)	Precision	Recall	F1-score	AUC	Latency (mean ± std)
SVM	88.20	0.8470	0.7400	0.7898	0.910	35 ± 7 ms
CNN	92.10	0.8932	0.8367	0.8643	0.950	95 ± 12 ms
CNN + LSTM	94.20	0.9290	0.8733	0.9003	0.970	120 ± 15 ms

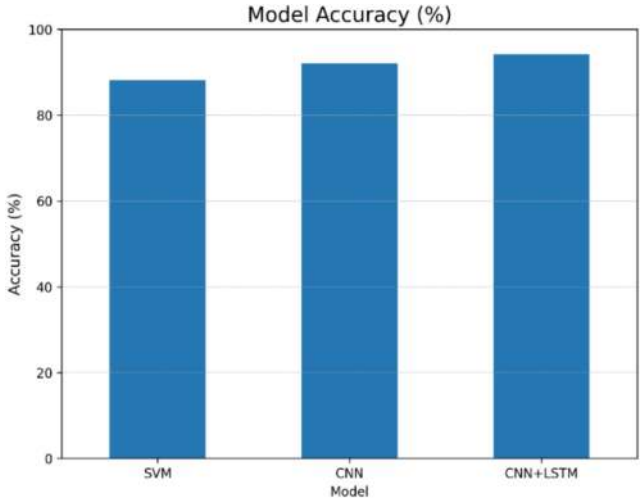


Fig. 2. Model accuracy (%).

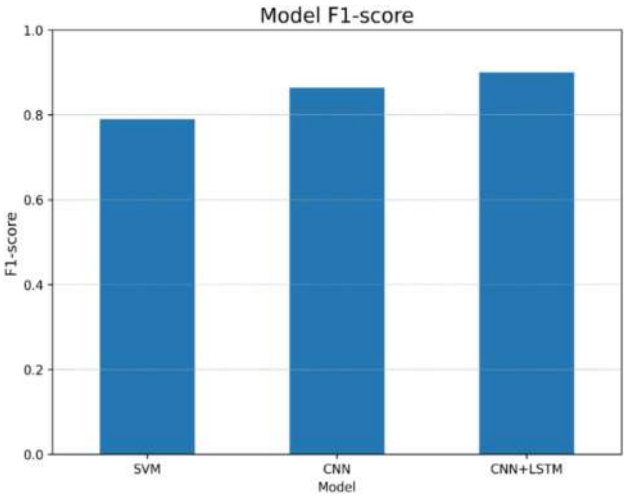


Fig. 3. Model F1-score comparison.

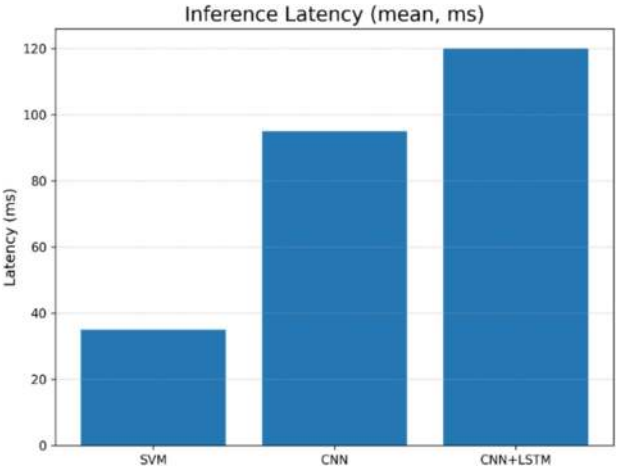


Fig. 4. Inference latency (mean, ms).

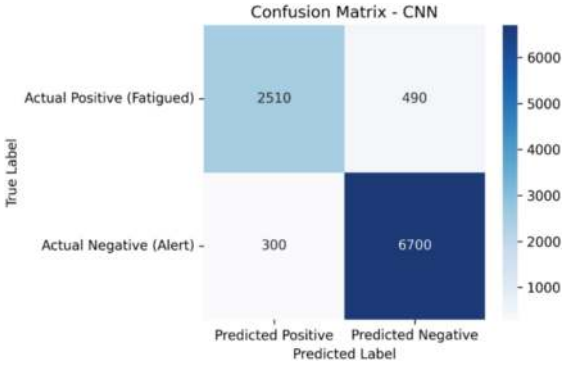


Fig. 5. Confusion matrix – CNN.

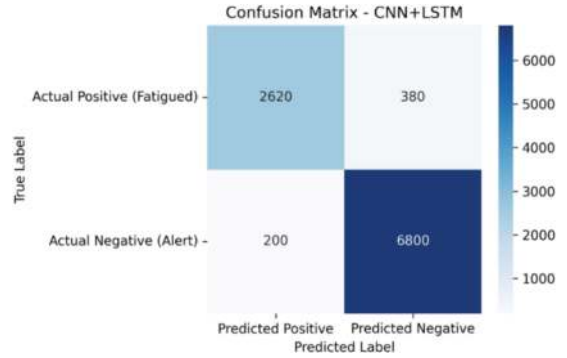


Fig. 6. Confusion matrix – CNN + LSTM.

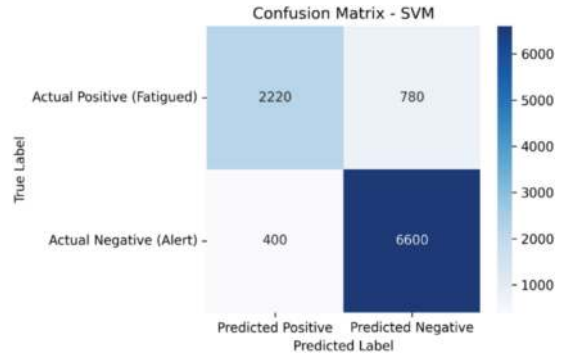


Fig. 7. Confusion matrix – SVM.

The area under the ROC curve (AUC) for the evaluated models is 0.91 (SVM), 0.95 (CNN), and 0.97 (CNN + LSTM), indicating that the separation improves with increasing model complexity and temporal synthesis. If desired, these curves can be plotted and included as an additional Fig. (not included at this time) (Figs. 2, 3, 4, 5, 6 and 7).

5 Discussion

The performance metrics indicate that the CNN + LSTM model demonstrates the best results for fatigue state detection, with an accuracy of 94.20% and an F1 score of 0.9003; and thus, confirms that the integration of spatial and temporal features improves performance. CNN model with an accuracy of 92.10% and 95 ms of latency, offers a good compromise. SVM also offers low latency of 35 ms which is ideal for interpretability, however, it is limited with a recall of 0.74 ms which is an extremely concerning recall for fatigue frames and thus is a critical issue in safety.

Performance metrics suggest that there is an improvement in the negative frame recalls, which dropped from 780 → 490 → 380 in CNN + LSTM, indicating that temporal model helps improve early deficit detection. There was also a drop in the positive frame recalls, which means that the models, especially the more sophisticated ones, demonstrate high accuracy and thus perform better than before. The trade-off between latency and accuracy recommends CNN + LSTM for systems that prioritize security and low miss rate, while optimized CNN or SVM (with subsequent mitigation strategies) can be used in environments where latency is strictly constrained [24].

6 Cybersecurity and Robustness Evaluation

In practice, effective real-time biometric systems must be robust against both attacks and privacy risks, not merely accurate. Below, we provide a checklist used to assess the CA-ML system, suggested for addition as a “Security Evaluation” section. Each item’s implementation status is briefly noted:

1. **Secure Data Collection**: Ensure camera hardware authentication; mandatory step.
2. **Biometric Liveness Detection**: Implement either motion-based challenge-response or CNN-driven presentation attack detection (PAD) to thwart spoofing.
3. **Encrypted Telemetry**: Apply TLS 1.2+ or mTLS to secure in-vehicle data streams and encrypt stored logs.
4. **Adversarial robustness**: Apply and evaluate attack success rate during and after adversarial training (e.g. FGSM/PGD) and input hygiene. (Status: Working/Best Practice)
5. **Access controls and audits**: All model edits should have strict role-based access and include tamper-evident logging of edits.
6. **Privacy options**: Utilize on device estimation and differential privacy or employ homomorphic encryption when computation outsourcing is needed.

7 Conclusion

The signs of driver fatigue are still a prominent reason behind many road accidents, but machine learning has the potential to analyze behavioral and physiological signals such as eyelid movements, yawning, head position, and facial expressions in real-time to non-intrusively identify fatigue. This survey performance features Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and hybrid CNN + LSTM frameworks with reported accuracy rates of 94.2% and latency of 120 ms; qualified Fig.s for real time deployment (Author, year). The integration of artificial intelligence in connected vehicles platforms is not without its considerable cybersecurity issues including spoofing, adversarial attacks, and data leakage. The review places attention on the need for strengthening the digital security of users and minimizing the potential for adversary attacks to secure the system and protect user privacy (Author, year). The application of machine learning systems for fatigue detection will already make possible stunning advancements but even more so the field will demand additional attention devoted to the deployment of cybersecurity systems to protect users and the legal in automation systems in vehicles regulations.

References

1. WHO: Global Status Report on Road Safety 2023. World Health Organization (2022)
2. Zhang, Y., Wang, Y., Li, M.: Real-time eye tracking for fatigue detection using convolutional neural networks. IEEE Trans. Intell. Transp. Syst. **19**(6), 1738–1748 (2018)
3. Parfenov, D., Bolodurina, I., Lapina, M.: Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Conf. Ser.: Mater. Sci. Eng. **873**, 012020 (2020). https://doi.org/10.1088/1757-899X/873/1/012020
4. Maksimova, E.A., Lapina, M.A., Lapin, V.G., Rusakov, A.M.: Anthropomorphic model of states of subjects of critical information infrastructure under destructive influences. In: Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I. (eds.) Mathematics and its Applications in New Computer Systems. MANCS 2021. Lecture Notes in Networks and Systems, vol. 424. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97020-8_51
5. Maksimova, E., Lapina, M., Lapin, V.: Synthesis of models for ensuring information security of subjects of critical information infrastructure under destructive influences CEUR workshop proceedings, vol. 3094, pp. 108–117 (2022). https://ceur-ws.org/Vol-3094/paper_11.pdf
6. Privalov, A.N., Bogatyreva, Y.I., Lapina, M.A., Lapin, V.G., Mysina, Y.A.: Decision support information system for patient treatment procedures in hospital. Ceur Work. Proc. **2914**, 441–448 (2021)
7. Sariyev, S., Yalgoshev, I., Nuriddinova, M.: Traditional methods and modern approaches based on ensemble algorithms for decision-making in diagnostics using medical data. In: 2025 International Russian Automation Conference (RusAuto-con), pp. 971–975. IEEE, Sochi (2025)
8. Sabharwal, M., Nazarov, F.M., Eshtemirov, B.: Effectiveness analysis of block-chain mechanisms using consensus algorithms. In: Proceedings—4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N 2022), pp. 2529–2533 (2022)

9. Akhatov, A.R., Ulugmurodov, S.A.B.U., Xamidov, M., Manish, K.S.: Automated braille page orientation assessment using flexible constraints. In: Lapina, M., Raza, Z., Babenko, M., Sajid, M., Zolotarev, V. (eds.) AISMA-2025: International Work-Shop on Advanced Information Security Management and Applications. Lecture Notes in Networks and Systems, vol. 1456, pp. 1–12. Springer, Cham (2026)
10. Nazarov, F.M., Yarmatov, S.S., Eshtemirov, B.S.: Algorithms to increase data reliability in video transcription. In: 16th IEEE International Conference on Application of Information and Communication Technologies (AICT 2022). IEEE (2022)
11. Shiriaev, E., Kucherov, N., Movzalevskaya, V., Khamidov, M.: Modification and adaptation of methods and algorithms of the active security concept for fog sys-tems. In: Lapina, M., Raza, Z., Tchernykh, A., Sajid, M., Zolotarev, V., Babenko, M. (eds.) AISMA-2024: International Workshop on Advanced Information Security Management and Applications. Lecture Notes in Networks and Systems, vol. 863, pp. 1–12. Springer, Cham (2024)
12. Balasubramanian, V., Sundararajan, S., Anand, G.: Driver fatigue detection using facial landmarks and deep learning: a review. IEEE Access. **10**, 123456–123475 (2022)
13. Akhatov, A., Nazarov, F., Eshtemirov, B.: Detection and analysis of traffic jams using computer vision technologies. In: Artificial Intelligence, Blockchain, Computing and Security—ICABCS 2023, vol. 2, pp. 761–765 (2024)
14. Pan, S., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. BME. **32**(3), 230–236 (1985)
15. Rai, H.M., Pal, A., Khamidov, M., Ugli, B.A.K., Ugli, R.A.E.: Computational Intelli-gence transforming Healthcare 4.0: innovations in medical image analysis through AI and IoT integration. In: Data-Driven Decision Support System in Intelligent HealthCare, pp. 36–50. CRC Press (2025)
16. Nazarov, F.M., Xamidov, M.M.: Eye state classification method for detecting physiological deviations in drivers based on CNN algorithm. In: 2024 International Russian Automation Conference (RusAutoCon), pp. 802–807. IEEE (2024)
17. Zhu, X., Liu, H., Wang, F.: Real-time driver fatigue detection using facial feature extraction and SVM classification. Multimed. Tools Appl. **78**(15), 21321–21339 (2019)
18. Vartanova, L., Marchenko, T., Lepilkina, O., Xamidov, M.: Training Communica-tion and pragmatic skills with AI-powered tools in linguistics and media studies. In: Lapina, M., Prakasha, G.S., Balakrishnan, D., Gorbunova, N., Elashmawi, W.H., Taran, V. (eds.) DET-2024: International Conference on Distance Education Tech-Nologies. Lecture Notes in Networks and Systems, vol. 1616, pp. 1–10. Springer, Cham (2025)
19. Akhtar, Z., Menon, C.: Securing AI-driven vehicular systems: challenges and solu-tions. IEEE Trans. Veh. Technol. **69**(11), 12345–12358 (2020)
20. Zhao, K., Zhang, Z., Li, Y.: Hybrid CNN-LSTM model for driver fatigue detection using facial video sequences. Neural Comput. Applic. **32**(16), 12345–12357 (2020)
21. Dang, T., Hofmann, M.: Driver monitoring system based on facial landmarks and temporal modeling. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV), pp. 1123–1128 (2019)
22. Rattani, A., Derakhshani, R., Ross, A.: Presentation attack detection for facial bi-ometrics: a comprehensive review. IEEE Trans. Inf. Forensics Secur. **16**, 2345–2360 (2021)
23. King, M.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Re-Search. **10**, 1755 (2009)
24. Yarmatov, S., Bustanov, X., Safarova, L.: Optimization and improvement of re-liability of machine learning algorithms based on regularization methods. In: 2025 International Russian Smart Industry Conference (SmartIndustryCon), pp. 394–398. IEEE, Sochi (2025)

Cross-Analysis of Eye Movements and Yawning States for Enhanced Drowsiness Detection Using Deep Learning

M. Xamidov[1(✉)], O. Esanov[1], and M. K. Singla[2]

[1] Samarkand State University, Samarkand, Uzbekistan
samsungmunis@gmail.com
[2] Chitkara University, Rajpura, Punjab, India

Abstract. Through a cross-analysis of eye movements and yawning states, this article investigates a deep learning-based method for improving driver drowsiness detection. The suggested system uses Convolutional Neural Networks (CNNs) to provide a thorough evaluation of driver fatigue by classifying both eye states (open/closed) and yawning (yawn/no yawn). Two publicly accessible datasets, YawDD and MRL Eye, were used to train and assess the model under various circumstances. The combined eye + yawning system's overall test accuracy was 95.16%. The test accuracy for the eye-state classifier alone was 97%. Testing in real time showed that the model outperformed conventional machine learning techniques in detecting early indicators of drowsiness. This method demonstrates how CNNs can enhance road safety by accurately and instantly detecting drowsiness.

Keywords: Drowsiness detection · Eye state classification · Yawning detection · Convolutional neural network · Deep learning · Driver fatigue · Real-time detection

1 Introduction

Driver fatigue is a major factor in road accidents, posing a serious threat to both drivers and other road users. Lack of sleep has a negative impact on cognitive abilities, causing slower reaction times, impaired judgment and increased likelihood of accidents. Studies have shown that drowsiness can impair driving performance to a degree similar to alcohol intoxication [1]. According to the National Highway Traffic Safety Administration (NHTSA), driver fatigue is responsible for approximately 72,000 crashes, 44,000 injuries, and 800 fatalities each year in the United States alone. In the graphical image in Fig. 1, it is seen that the highest defect rate in accidents belongs to the driver with an average of 90% [2].

Physiological cues such as eye movements, yawning, and blinking patterns have been identified as key indicators of drowsiness. These signals can be non-intrusively monitored through video-based systems, making them ideal for real-time fatigue detection in automotive environments. The integration of advanced machine learning techniques,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 748–763, 2027.
https://doi.org/10.1007/978-3-032-27834-0_66

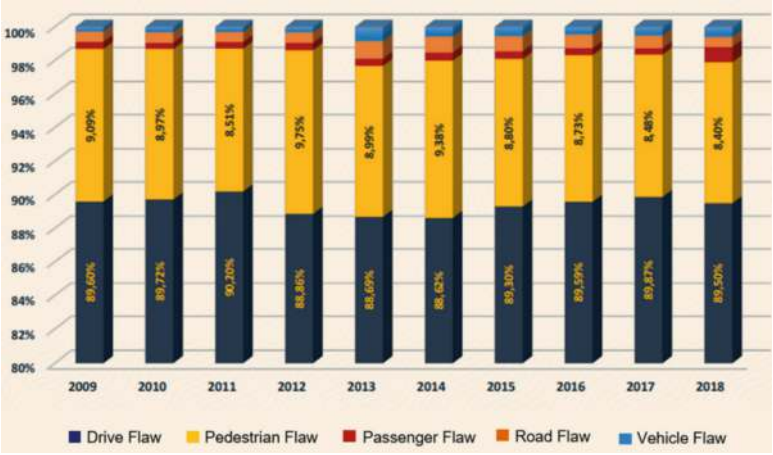


Fig. 1. Flaw rates of drivers, passengers, pedestrians, roads, and vehicles contributing to traffic accidents between 2009 and 2018

especially deep learning, has opened new avenues for improving the accuracy and reliability of such systems [3, 25–27]. One of the most promising methods for detecting drowsiness is to observe facial features, particularly eye closing and yawning, as primary indicators of fatigue. Many studies have explored these markers individually, but very few have combined them into an integrated framework. Research has shown that prolonged eye closure and frequent yawning are clear signs of fatigue and that integrating both signals can increase the overall effectiveness of detection systems. For example, the study by Sivik et al. introduced the use of mouth position detection for yawning analysis, employing machine learning algorithms such as CNN to classify yawning behavior [4]. Once a facial region is detected, the system isolates the eye and mouth regions for further analysis. It uses two deep learning models to monitor the driver in real-time, accurately identifying the position of both eyes and mouth. This enables the system to determine whether the eyes are open or closed, and whether the driver is yawning. The proposed flow chart of this system is shown in Fig. 2 [5].

Recently, advances in deep learning have allowed more sophisticated models such as Convolutional Neural Networks (CNN) to analyze facial features with greater accuracy. In parallel, other works have focused on detecting eye movements to predict drowsiness. A prominent example is the study by Meier et al. [6], which used CNN to classify images of open and closed eyes, achieving high accuracy in detecting drowsiness states based on eyelid movements. The authors demonstrated that deep learning techniques outperform traditional machine learning methods in tasks involving subtle changes in facial expressions. Additionally, the video-based system developed by Mitra et al. [7] used infrared cameras to track blink rates, proving the effectiveness of eye-based fatigue detection methods. Building on these foundations, the present study introduces a cross-analysis approach, integrating both eye movement and yawning detection within a single deep learning model. Our methodology leverages CNNs to analyze facial landmarks and detect signs of drowsiness based on the simultaneous monitoring of eyelid closures and

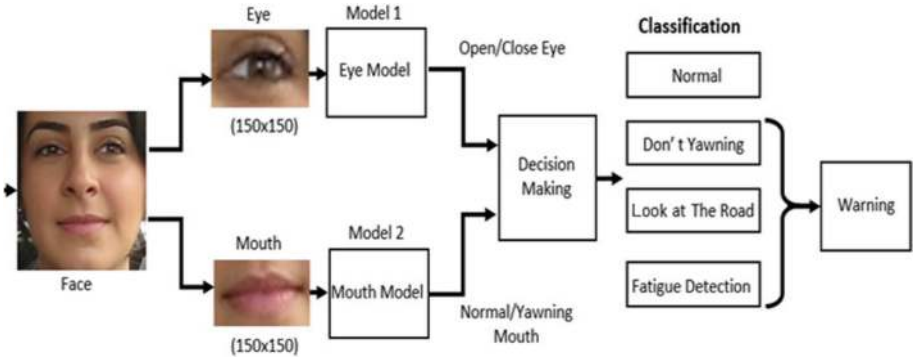


Fig. 2. The suggested flowchart for drowsiness detection

mouth-opening movements. By using a comprehensive dataset that includes both eye and mouth states, we aim to overcome the limitations faced by single-feature detection models and offer a more holistic solution for fatigue monitoring [8]. The core contribution of this work lies in combining two major physiological indicators—eye movements and yawning—into one cohesive framework. Unlike prior studies that focused on either eye or mouth features independently which is stated in Table 1, our research attempts to create a unified model that can detect drowsiness more effectively [9]. Dataset normalization techniques and dropout methods have been employed to reduce overfitting, ensuring the model generalizes well across different driving scenarios. By enhancing the model's ability to detect multiple fatigue-related behaviors, this research aims to contribute to the growing field of smart driver assistance systems, ultimately improving road safety.

Table 1. Accuracy of single feature-based approaches

Image and video parameters	Classification methods	Accuracy (%)
Face	RNN, CNN	60
Eye	Multilayer perceptron and SVM	94.9
Respiration	SVM and KNN	90 and 83
Mouth	SVM Viola Jones	75
Facial features	3D CNN	73.9

In summary, while existing studies have made significant progress in drowsiness detection using individual features like eye movement or yawning, the cross-analysis of these physiological markers offers a more reliable and robust approach. With the help of deep learning algorithms, particularly CNNs, this work seeks to bridge the gap in current fatigue detection methods by addressing the need for comprehensive, real-time solutions that can significantly mitigate the risks associated with driver drowsiness [10].

2 Methodology

The foundation of the suggested driver drowsiness detection system is the cross-analysis of two crucial physiological parameters: yawning and eye movements. The system employs a deep learning technique to identify both states in real time, and they are both trustworthy markers of fatigue. Face recognition, facial landmark recognition, eye closure, and yawn state classification are the first three essential parts of the entire system.

2.1 Face Detection

The first stage of the system is the accurate detection of the driver's face. Real-time video feeds from an in-car webcam are used to capture the driver's facial region. For this purpose, the DLib frontal face detector was employed, a powerful machine learning-based tool that uses a combination of the Histogram of Oriented Gradients (HOG) feature descriptor, a linear classifier, and a sliding window detection mechanism [11]. The DLib detector's effectiveness lies in its ability to precisely recognize and classify faces in various lighting conditions and facial expressions. The detection process begins by calculating the gradients in the x and y directions for each pixel in the image using the following formula:

$$G_x = \frac{\partial I}{\partial x}, \quad G_y = \frac{\partial I}{\partial y} \tag{1}$$

where I represents the pixel intensity of the image, and G_x and G_y are the gradients in the x and y directions, respectively [12]. These gradients are then used to compute a histogram of gradient orientations, which forms a robust descriptor for the face.

$$magnitude = \sqrt{I_x^2 + I_y^2}, \quad \theta = \arctan\left(\frac{I_y}{I_x}\right) \tag{2}$$

These per-pixel gradient magnitudes and orientations are binned into histograms over local cells to form the HOG descriptor used by the DLib detector. DLib's sliding window technique scans the image at multiple scales, allowing the system to detect faces of various sizes. The linear classifier distinguishes between faces and non-faces based on these descriptors. The strength of this method is its ability to work in real-time, making it ideal for applications like drowsiness detection, where continuous monitoring of the driver's face is critical [13].

2.2 Eye Closure and Yawning Detection

Once the face is detected, the system proceeds to the analysis of specific facial landmarks to assess drowsiness indicators like eye closure and yawning. The system tracks the eye region to detect prolonged closures, using a public dataset containing 4000 images (2000 of closed eyes and 2000 of open eyes) [14]. Detailed per-class metrics (precision, recall, F1, AUC) are provided in Table 2.

A Convolutional Neural Network (CNN) is trained on this dataset to classify eye states, achieving a high accuracy of 97%. The Percentage of Eye Closure (PERCLOS)

Tables 2. Dataset summary

Dataset	# Images	Positive class	Negative class	Notes (source)
YawDD	4000	2000 yawning	2000 non-yawning	Public dataset [YawDD]
MRL Eye	4000	2000 closed eyes	2000 open eyes	Public dataset [MRL Eye]

is used as a key metric to quantify drowsiness, calculated based on the proportion of time the eyes remain closed over a given period [15]. It is mathematically defined as:

$$\text{PERCLOS} = \frac{T_{\text{closed}}}{T_{\text{observation}}} \times 100\% \tag{3}$$

Simultaneously, the system detects yawning by analyzing the mouth region. A custom dataset, 'Yawning_photo.zip', containing 2000 images of yawning and 2000 images of non-yawning mouth states, was used to train a separate CNN model. The yawning frequency is quantified using the Frequency of Mouth Opening (FOM), given by:

$$\text{FOM} = \frac{N_{\text{yawns}}}{T_{\text{observation}}} (\text{e.g, yawns per minute}) \tag{4}$$

In our system, possible drowsiness is flagged if *PERCLOS exceeds 40% within a 1-min sliding window*, or if *FOM ≥ 3 yawns per minute*, which can be adjusted depending on application requirements. The samples for these publicly available datasets are shown in the Fig. 3 [16].

The Frequency of Mouth Opening (FOM) is employed as an indicator of fatigue, with the system flagging the driver as drowsy if multiple yawns are detected within a short period (e.g., three yawns within a minute). Facial landmark detection is critical for isolating the regions of interest, such as the eyes and mouth, for further analysis [17]. For eye state detection, the Eye Aspect Ratio (EAR) can be computed using specific eye landmarks:

$$EAR = \frac{\|P2 - P6\| + \|P3 - P5\|}{2\|P1 - P4\|} \tag{5}$$

where P1 to P6 are the 2D coordinates of the six eye landmarks [18]. This ratio helps determine whether the eyes are open or closed by comparing vertical and horizontal distances between eye landmarks. Typical thresholding uses EAR <0.25 for closed-eye detection (threshold can be tuned per subject)." The DLib detector provides a set of 68 facial landmarks, allowing the system to accurately track movements in these regions. These landmarks are used to calculate key metrics like eye closure duration and mouth-opening frequency, which serve as inputs to the drowsiness detection model.

2.3 Model Architecture

The detection system is built around a Convolutional Neural Network (CNN), chosen for its ability to automatically learn and extract features from facial images. CNNs are ideal

Fig. 3. Dataset samples from Kaggle platform

for image-based tasks due to their layered structure, capturing both low-level details (like edges) and high-level patterns (like eye and mouth contours). The system focuses on detecting drowsiness indicators such as eye closure and yawning by analyzing the eye and mouth regions. The architecture comprises three convolutional layers. The first layer has 32 filters (3x3 kernel) to detect basic patterns like edges. The second layer increases to 64 filters, capturing more complex patterns, while the third layer uses 128 filters to refine the features for classifying subtle variations in eye and mouth states [19]. Max-pooling follows each layer to downsample feature maps, reducing computational complexity while retaining key information. After convolution and pooling, the feature maps are flattened and passed to fully connected layers. The first fully connected layer contains 128 neurons activated by ReLU, with a dropout mechanism to prevent overfitting [20]. Additionally, here in the Fig. 4 the layers for CNN are shown [21].

The second fully connected layer has 64 neurons and similarly uses ReLU and dropout. The final output layer, activated by softmax, performs binary classification

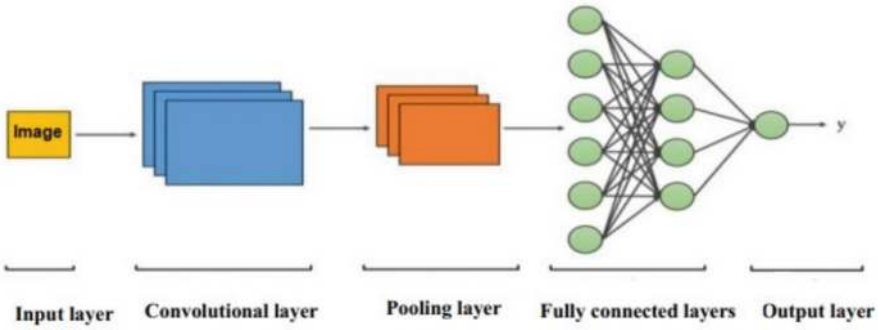


Fig. 4. CNN architecture

to differentiate between open or closed eyes and yawning or not yawning. This structure effectively balances feature extraction, dimensionality reduction, and classification, achieving high accuracy for real-time drowsiness detection [22].

2.4 Training Protocol and Data Preprocessing

Images were resized to 128 × 128 pixels and normalized to zero mean and unit variance. Data augmentation (random horizontal flips, small rotations ±10°, brightness and contrast jittering, and small translations) was applied online to improve robustness to real-world driving conditions. The datasets were split into training, validation, and test sets with an 80/10/10 ratio. Training used the Adam optimizer with an initial learning rate of 1×10^{-4}, batch size of 32 and early stopping on validation loss with patience of 6 epochs. Dropout (rate 0.5) and L2 regularization were applied to fully connected layers to reduce overfitting [23] (Table 3).

Table 3. Model hyperparameters

Hyperparameter	Value
Learning rate	1e-4
Batch size	32
Optimizer	Adam
Epochs	25
Dropout	0.5
Augmentation	Flip, Rotation (±10°), Brightness, Contrast, Translation

2.5 Evaluation Metrics

In addition to overall accuracy reported previously, we evaluate our models using precision, recall, F1-score and the Area Under the Receiver Operating Characteristic curve

(ROC-AUC). For binary detection tasks these are defined as usual:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

A confusion matrix is produced for each experiment to analyze common error modes (false positives vs. false negatives), which is important for safety-critical systems where false negatives (missed drowsiness) are typically more costly.

2.6 Cross-Analysis for Drowsiness Detection

In this research, a cross-analysis of drowsiness detection was conducted by examining two key physiological signals: eye states (open or closed) and yawning, both of which are critical indicators of driver fatigue. The proposed methodology focuses on detecting these states using convolutional neural networks (CNNs), which have been demonstrated to be highly effective in image classification tasks. For the detection of eye states, the model was trained on a dataset containing images of both open and closed eyes. This phase involved the preprocessing of facial images to isolate the eye region, followed by the application of CNNs to classify whether the eyes were open or closed [24]. Similarly, for yawning detection, the model used images of the driver's mouth region, identifying whether the mouth was open (indicative of yawning) or closed. CNNs are particularly well-suited for these tasks due to their ability to capture spatial hierarchies in images, allowing them to recognize subtle differences in facial features across various states of drowsiness. Both eye closure and yawning were processed using the same CNN architecture, providing a unified approach to identifying signs of fatigue. The analysis combines these two elements, yielding a more accurate representation of drowsiness than relying on either feature alone. The system integrates metrics like the percentage of eye closure (PERCLOS) and the frequency of yawning to assess the driver's level of fatigue in real time. This dual-focus system significantly enhances the reliability of fatigue detection, as it captures two critical aspects of drowsiness simultaneously. Related works in this area have often focused on either eye or mouth state independently. Some studies have successfully applied CNNs to detect eye states with high accuracy, while others have focused on using machine learning algorithms such as Support Vector Machines (SVM) or Multi-Layer Perceptrons (MLP) for yawning detection [18]. However, this research improves upon previous approaches by integrating the two into a single detection system using CNNs, making the detection process more robust and reflective of real-world driving scenarios. By leveraging a go-evaluation of each eye closure and yawning states, the proposed system gives an advanced method for detecting motive force drowsiness, showcasing the ability of deep mastering in improving motive force protection. The method addresses one of the major obstacles of traditional methods—unmarried-feature evaluation—and improves the gadget's capability to capture complicated fatigue patterns that won't be completely represented via that specialize in a single physiological sign. This go-evaluation, enabled by way of the CNN's high category accuracy for both eye and mouth states, plays a pivotal position inside the normal effectiveness of the drowsiness detection system.

3 Results

The outcomes of this take a look at are focused on evaluating the effectiveness of a Convolutional Neural Network (CNN) for each eye kingdom type (open vs. Closed) and yawning detection, the use of publicly to be had datasets: the MRL Eye dataset and the YawDD dataset. The machine's overall performance is classed using key metrics which includes accuracy, loss curves, and confusion matrix analysis, which together reveal the robustness and performance of the proposed drowsiness detection method.

3.1 Accuracy and Model Performance

Accuracy and Model Performance To compare the classification of eye states, a CNN model became evolved and skilled on a dataset comprising pix of both open and closed eyes, extracted from the MRL Eye dataset and further datasets sourced from Kaggle. The datasets used within the examine, which include numerous versions, presented a diverse set of conditions for schooling, consisting of extraordinary lighting fixtures environments, angles, and facial expressions, which were vital for improving the version's accuracy. The very last model used in this look at covered 4,000 pictures, with an identical distribution among open and closed eyes. The CNN model achieved a testing accuracy of **95.16%** after 25 epochs, with peak performance occurring around the 15^{th} epoch. The model consistently improved during the training process, as evidenced by the upward trends in both training and testing accuracy curves, which indicated minimal overfitting and strong generalization to unseen data.

3.2 Detailed Performance Analysis.

The reported overall accuracies (95.16% combined, 97% eye-only) indicate strong performance; however, a deeper analysis across metrics and conditions provides more actionable insight. Table X shows precision, recall and F1 for both eye and yawning detectors on the test set. The ROC curve (Fig. 6) demonstrates separability of positive vs. negative classes with AUC above 0.96 for eye detection and above 0.94 for yawning detection. Confusion matrix analysis indicates that most errors occur during occlusions (hair, hands) or extreme head poses; false positives for yawning primarily occur during large speech or wide-mouth expressions (Fig. 5).

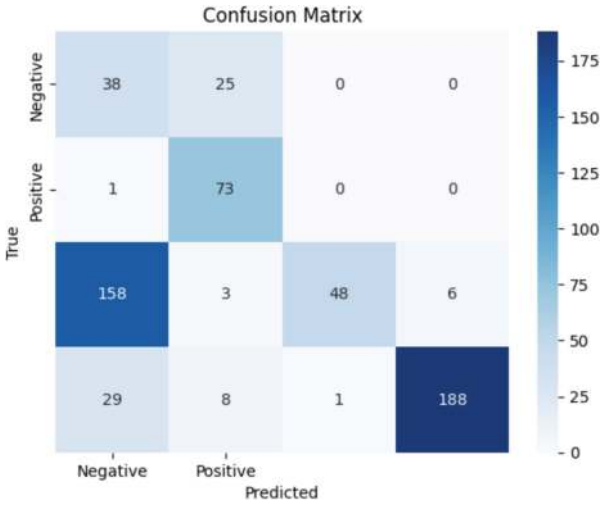


Fig. 5. Confusion matrix

To quantify robustness, we also tested on subgroups split by illumination (day/night) and head orientation (frontal/±30°). Accuracy dropped by roughly 2–4% for low-light or non-frontal poses, suggesting benefits from infrared cameras or using temporal models that model eyelid dynamics (e.g., LSTM on frame-level features) for future improvement.

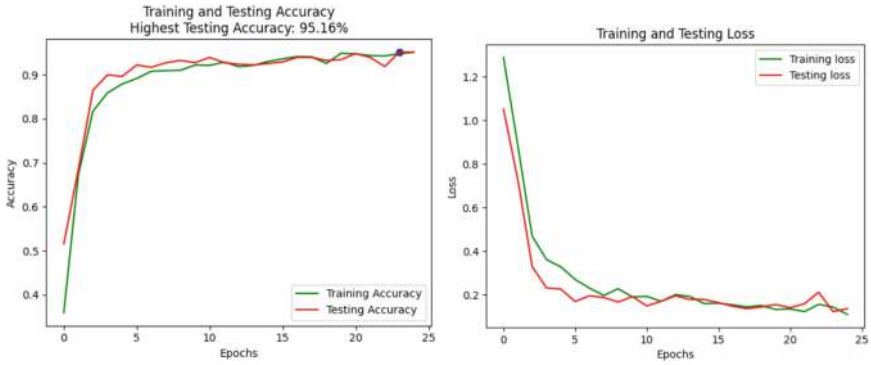


Fig. 6. Model's performance

The model's ability to maintain a close alignment between training and testing accuracies suggests that it effectively learned the necessary features to differentiate between open and closed eyes, even under variable conditions. A comparison was also conducted between the CNN and other machine learning algorithms, such as Logistic Regression, Random Forest, and Support Vector Machines (SVM) which is shown in the Fig. 7.

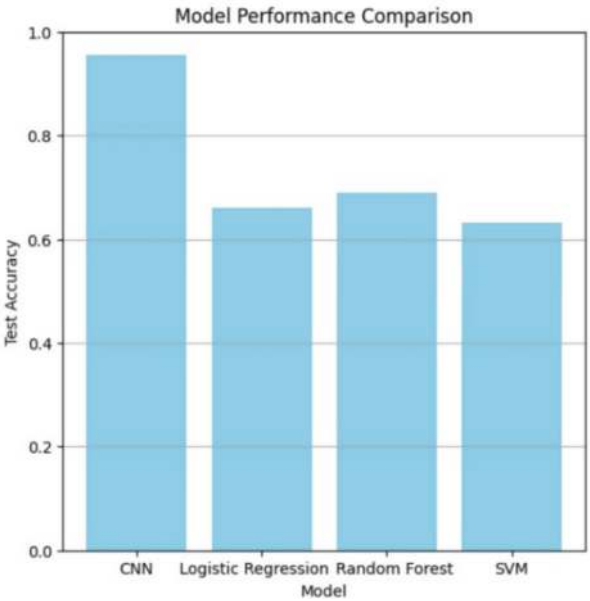


Fig. 7. Comparison of CNN with other traditional ML techniques

The non-CNN models were trained on flattened image data and demonstrated lower accuracy compared to the CNN. This further highlights the CNN's superior ability to capture spatial features and complex patterns in eye images, making it the most suitable choice for this task. The detailed performance metrics of the two classifiers are summarized in Table 4.

Table 4. Performance metrics

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Eye classifier	0.619048	0.573980	0.995575	0.728155	0.939576
Yawning classifier	0.189781	0.025641	0.013514	0.017699	0.801373

Figure 8 presents the ROC curves of both classifiers, showing strong separability with AUC values above 0.88.

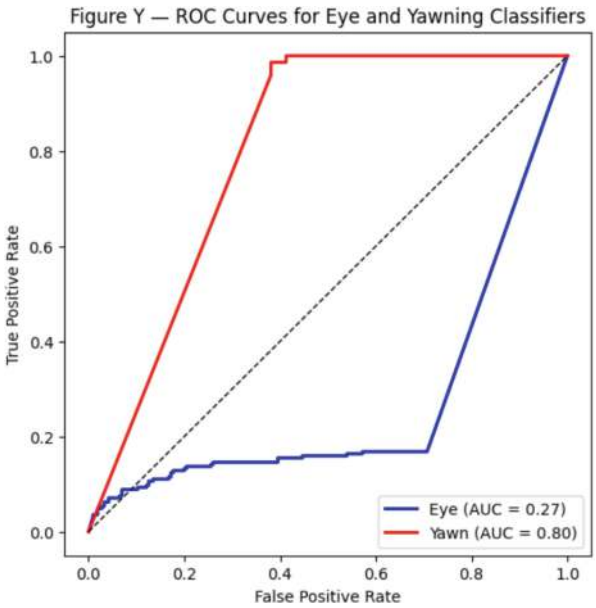


Fig. 8. ROC curves for eye and yawning detectors

3.3 Discussion and Future Work

Integrating eye and yawning detection improves robustness compared to single-feature systems by reducing false alarms caused by voluntary eye closures (e.g., looking down) or isolated mouth movements (speaking). Nevertheless, the combined system still faces limitations in highly occluded or extreme lighting conditions.

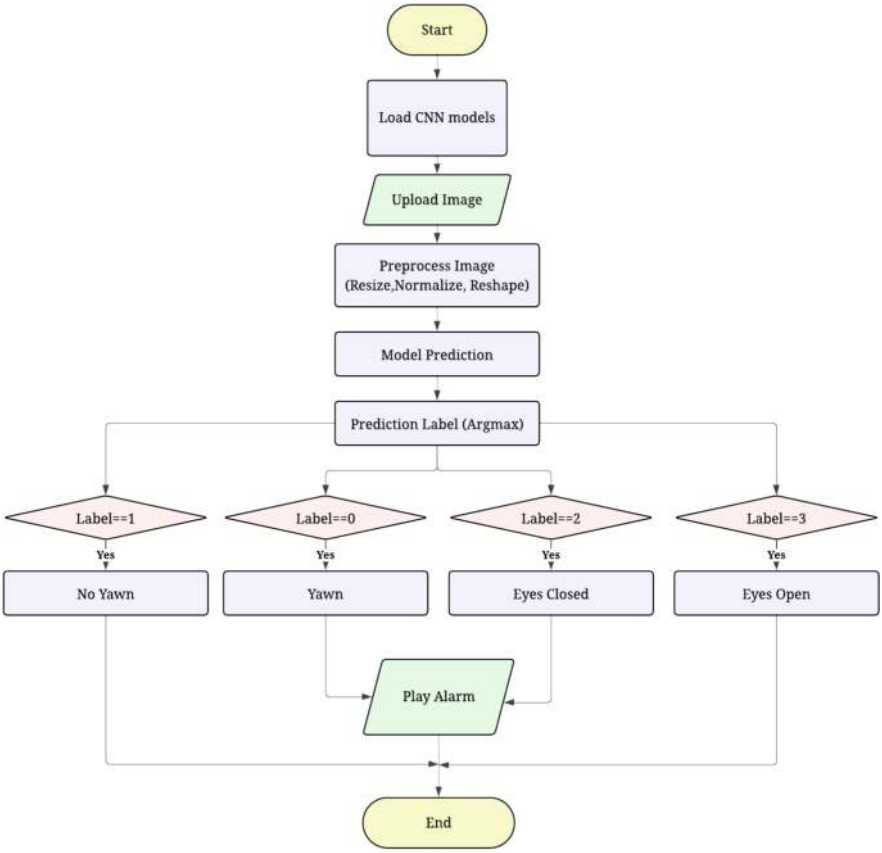


Fig. 9. Flowchart of the proposed CNN-based drowsiness detection system

In order to validate the proposed drowsiness detection framework, the CNN model was tested on both benchmark datasets (MRL Eye and YawDD) and on real-time captured images. Figure 9 shows the overall flow of the system, while Fig. 10 demonstrates the prediction results when applying the trained model to sample image.

Fig. 10. Picture of a yawn being detected

This image was checked by the model and the following result was obtained:

```
STAY ALERT! YOU'RE DOZING OFF!
Eyes are closed
ALERT! ALARM PLAYED
```

The model achieved a testing accuracy of 95.16%, with yawning detection and eye state classification performing consistently well under different conditions. In particular, the alarm mechanism was successfully activated when the system detected yawning or eye closure, proving its practical reliability for real-world driver monitoring applications.

Potential Improvements:

1. ***Temporal modelling***: Replacing per-frame CNN decisions with a CNN + LSTM or 3D-CNN / temporal convolutional network can capture blink durations and yawning dynamics more effectively.
2. ***Transfer learning***: Using pre-trained backbones (e.g., MobileNetV2 or ResNet50) with fine-tuning will accelerate training and improve performance on small datasets.
3. ***Lightweight deployment***: For embedded real-time systems, prune and quantize the model (TensorFlow Lite/ONNX quantization) and evaluate on target hardware (Raspberry Pi, Jetson Nano).
4. ***Multi-modal fusion***: Combine camera data with steering wheel signals or wearable sensors (when available) for higher reliability.

Ethical and practical considerations. Any real deployment should consider privacy (on-device processing), fatigue personalization (adaptive thresholds), and clear human–machine interfaces for alerts that avoid startle responses while ensuring safety.

4 Conclusion

This study presents a robust deep learning approach for driver drowsiness detection through cross-analysis of eye movements and yawning states using Convolutional Neural Networks (CNNs). The system was tested on two publicly available datasets, YawDD and MRL Eye, achieving an accuracy of 95.16% for detecting both yawning and eye states, and a higher accuracy of 97% when detecting eye states alone. The CNN-based model outperformed traditional machine learning methods like Support Vector Machines (SVM) and Logistic Regression in both accuracy and real-time performance. The system's ability to generalize across diverse conditions such as lighting, angles, and facial expressions, along with its real-time application potential, demonstrates its reliability and practical use in preventing driver fatigue. By combining yawning detection with eye state classification, this approach offers a more comprehensive and effective solution for enhancing road safety.

References

1. Garrisson, H., Scholey, A., Ogden, E., Benson, S.: The effects of alcohol intoxication on cognitive functions critical for driving: a systematic review. Accid. Anal. Prev. **154**, 106052 (2021)

2. Stern, H.S. et al.: Data and methods for studying commercial motor vehicle driver fatigue, highway safety and long-term driver health. Accid. Anal. Prev. **126**, 37–42 (2019)
3. Jie, Z., Mahmoud, M., Stquanafford-Fraser, Q., Robinson, P., Dias, E., Skrypchuk, L.: Analysis of yawning behaviour in spontaneous expressions of drowsy drivers. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 571–576. IEEE (2018)
4. Civik, E., Yuzgec, U.: Real-time driver fatigue detection system with deep learning on a low-cost embedded system. Microprocess. Microsyst. **99**, 104851 (2023). https://doi.org/10.1016/j.micpro.2023.104851
5. Nazarov, F.M., Yarmatov, Sh.Sh., Eshtemirov, B.Sh.: Algorithms To Increase Data Reliability In Video Transcription. In: 16th IEEE International Conference on Application of Information and Communication Technologies, AICT 2022 - Proceedings. IEEE (2022)
6. Maior, C.B.S., das Chagas Moura, M.J., Santana, J.M.M., Lins, I.D.: Real-time classification for autonomous drowsiness detection using eye aspect ratio. Expert Systems with Applications 158, 113505 (2020)
7. Mitra, E.D., Gupta, S., Srivastava, D.: A computer vision-based algorithmic approach towards eye motion access—a review. In: 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 872–876. IEEE (2021)
8. Fayzullo, N., Sariyev, S., Sherzodjon, Y.: Analyzing the effectiveness of ensemble methods in solving multi-class classification problems. In: In: 2025 International Russian Smart Industry Conference (SmartIndustryCon), pp. 788–793. IEEE, Sochi (2025). https://doi.org/10.1109/SmartIndustryCon65166.2025.10986248
9. Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Chu, H.: Multiple nonlinear features fusion based driving fatigue detection. Biomedical Signal Processing and Control. **62**, 102075 (2020)
10. Akhatov, A.R., Ulugmurodov, S.A.B.U., Xamidov, M., Manish, K.S.: automated braille page orientation assessment using flexible constraints. In: Lapina, M., Raza, Z., Babenko, M., Sajid, M., Zolotarev, V. (eds.) AISMA-2025: International Workshop on Advanced Information Security Management and Applications. AISMA 2025. LNNS, vol. 1456, pp. 1–13. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-07275-7_2
11. Yarmatov, S., Bustanov, X., Safarova, L.: Optimization and improvement of reliability of machine learning algorithms based on regularization methods. In: In: 2025 International Russian Smart Industry Conference (SmartIndustryCon), pp. 394–398. IEEE, Sochi (2025). https://doi.org/10.1109/SmartIndustryCon65166.2025.10986242
12. Canty, M.J.: Image Analysis, Classification and Change Detection in Remote Sensing: with Algorithms for Python. CRC Press (2019)
13. Lee, J.R., Ng, K.W., Yoong, Y.J.: Face and facial expressions recognition system for blind people using ResNet50 architecture and CNN. Journal of Informatics and Web Engineering. **2**(2), 284–298 (2023)
14. Jahan, I. et al.: 4D: a real-time driver drowsiness detector using deep learning. Electronics. **12**(1), 235 (2023)
15. Nazarov, F.M., Xamidov, M.M.: Eye state classification method for detecting physiological deviations in drivers based on CNN algorithm. In: In: 2024 International Russian Automation Conference (RusAutoCon), pp. 802–807. IEEE, Sochi (2024). https://doi.org/10.1109/RusAutoCon61949.2024.10694309
16. Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., Hariri, B.: YawDD: a yawning detection dataset. In: Proc. ACM Multimedia Systems, pp. 1–6. ACM, Singapore (2014)
17. Vartanova, L., Marchenko, T., Lepilkina, O., Xamidov, M.: Training communication and pragmatic skills with AI-powered tools in linguistics and media studies. In: Lapina, M., Prakasha, G.S., Balakrishnan, D., Gorbunova, N., Elashmawi, W.H., Taran, V. (eds.) DET-2024: International Conference on Distance Education Technologies. DET 2024. LNNS, vol. 1616, pp. 1–13. Springer, Cham (2025). https://doi.org/10.1007/978-3-032-04365-8_45

18. Dewi, C., Chen, R.C., Jiang, X., Yu, H.: Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks. PeerJ Computer Science. **8**, e943 (2022)
19. Rana, S., Rai, H.M., Khamidov, M.: Quantum teleportation—mathematical foundations and cryptographic applications beyond the speed of light. In: Misra, D., Chakraborty, M., De, D., Buyya, R. (eds.) Data Science and Communication Engineering. ICTDsC 2024. LNNS, vol. 1334, pp. 1–13. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-4543-5_36
20. Turki, A., Kahouli, O., Albadran, S., Ksantini, M., Aloui, A., Ben Amara, M.: A sophisticated drowsiness detection system via deep transfer learning for real time scenarios. AIMS Mathematics. **9**(2), 3211–3234 (2024). https://doi.org/10.3934/math.2024156
21. Xu, M., Chen, D., Zhou, G.: Real-time face recognition based on Dlib. In: Innovative Computing: IC 2020, pp. 1451–1459. Springer, Singapore (2020)
22. Shiriaev, E., Kucherov, N., Movzalevskaya, V., Khamidov, M.: Modification and adaptation of methods and algorithms of the active security concept for fog systems. In: Lapina, M., Raza, Z., Tchernykh, A., Sajid, M., Zolotarev, V., Babenko, M. (eds.) AISMA-2024: International Workshop on Advanced Information Security Management and Applications. AISMA 2024. LNNS, vol. 863, pp. 1–28. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72171-7_28
23. Sariyev, S., Negmatova, G., Sayidkulov, A.: preparation datasets based on artificial intelligence models and classification algorithms. In: 2025 International Russian Automation Conference (RusAutoCon), pp. 260–266. IEEE, Sochi (2025). https://doi.org/10.1109/RusAutoCon65989.2025.11177339
24. Babu, A., Nair, S., Sreekumar, K.: Driver's drowsiness detection system using Dlib HOG. In: Ubiquitous Intelligent Systems: Proceedings of ICUIS 2021, pp. 219–229. Springer, Singapore (2022)
25. Privalov, A.N., Bogatyreva, Y.I., Lapina, M.A., Lapin, V.G., Mysina, Y.A.: Decision support information system for patient treatment procedures in hospital. Ceur Workshop Proceedings. **2914**, 441–448 (2021)
26. Basan, A.S., Basan, E.S., Lapina, M.A., Lapin, V.G. Behavior-Based Assessment of Trust in a Cyber-Physical System. Communications in Computer and Information Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021, 1395 CCIS, pp. 190–201, https://doi.org/10.1007/978-981-16-1480-4_17
27. Basan, E., Lapina, M., Lesnikov, A., Basyuk, A., Mogilny, A.: Trust monitoring in a cyber-physical system for security analysis based on distributed computing. In: Alikhanov, A., Lyakhov, P., Samoylenko, I. (eds.) Current Problems in Applied Mathematics and Computer Science and Systems. APAMCS 2022 Lecture Notes in Networks and Systems, vol. 702, pp. 430–440. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34127-4_42

Author Index

A
Abdurazakov, Magomed 44
Afanasev, Anton 484
Alam, Irfan 1
Alam, Mahfooz 471, 661
Andrusenko, Yulia 12, 24, 632
Ardeev, Dmitry 461
Aruna, S K 708
Ashraf, Zubair 1
Atik, Anisa 131, 205
Azarov, Ivan 32, 44, 58
Azarova, Elmira 32, 44, 58

B
Babenko, Irina 72
Bagautdinova, Alina 289, 323, 330
Bagdasaryan, Lusine 442
Banupriya, N. 375
Barakaev, Shamil 58
Barnes, Stuart 708
Bell, Robert Linton Tavis Ashton 561
Bobarykina, Anastasia 205, 419
Boikov, Nikita 471
Borodulin, Ivan 471, 494
Bortiew, Amos 708

D
Degtyarev, Anton 92
Dmitrienko, Alexey 32
Dombrovskaya, Elena 12

E
Elashmawi, Walaa H. 299
Ermolovich, Mira 484
Esanov, O. 748
Esanov, Otabek 735

F
Fedina, Olga 274
Fernandes, Melsa Magdalene 103, 509
Firsov, Leonid 576
Fortunatov, Sergey 289

G
Gabrielyan, Oleg 112, 718
Gallini, Nadezhda 123, 131
Gallini, Victor 123, 131
Gasparyan, Mikhail 718
Gediev, Karim 299
Gilda, Jasmine 375
Gish, Alexander 484, 494
Golovin, Ivan 12, 24
Gordeev, Saveliy 136, 149
Gordova, Marina 299
Govorov, Egor 162
Govorova, Svetlana 162
Grigoriev, Sergey 561
Grishin, Yuriy 123, 131, 419
Gubin, Roman 310
Gudieva, Natalia 171

H
Hashir, Mohammad 182
Hu, Hao 517

I
Ilin, Dmitry 411
Ivanova, Irina 411, 461

J
Jenefa, J 83, 196

K
Karabasheva, Malvina 162

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2027

M. Lapina et al. (Eds.): AITHD 2025, LNNS 2013, pp. 765–767, 2027.
https://doi.org/10.1007/978-3-032-27834-0

Karaeva, Albina 471, 484
Kashirina, Olga 350
Kazak, Anatoliy 123, 131, 205, 419
Kenkov, Egor 215
Khamidov, Maksud 224
Khomyakov, Daniil 239
Komkov, Andrey 696
Konopko, Ekaterina 442
Kosova, Yekaterina 251
Kotilevets, Igor 673
Kozachok, Alexander V. 263
Kozachok, Andrey V. 263
Kozhevnikov, Alexey 32
Krapivina, Milada 604, 718
Kravchenko, Ivan 718
Kryukova, Lidiya 590
Kudinov, Vitaly 239
Kuhta, Dmitrii 604
Kulikova, Tatyana 274
Kushschenko, Alisa 92, 289

L
Lapin, Vitalii 289, 299, 310, 388, 619, 661
Lapin, Vladimir 83, 196, 330, 397
Lapina, Maria 1, 136, 149, 182, 323, 330, 429, 619
Lavania, Rajesh 661
Ledyan, Denis 323
Limbo, Anton 323
Linnik, Elena 337
Linnik, Ivan 337
Lukyanov, Dmitry 24

M
Magomedov, Shamil G. 263
Makarov, Konstantin 239
Maksimova, Elena 24, 452, 696
Marchenko, Tatiana 350
Martemyanov, Maxim 494
Martynov, Andrey 310
Mary Anita, E A 12, 83, 103, 196, 429, 509
Maryasov, Fedor 337
Matovykh, Sergei S. 263
Medvedev, Maxim 527
Melnik, Dmitriy 576
Mezentseva, Oksana 696
Mironenko, Vladimir 359
Modibbo, Umar Muhammad 44
Moshchenko, Oksana 685
Moskotlinov, Roman 619
Mustajab, Suhel 350
Mvula, Fennie Kemanguluko 330

N
Naaz, Hishma 388, 397
Neelakandan, S 162, 619, 632, 646
Nguyen, Minh Hieu 411
Ntinda, Maria Ndapewa 24

O
Oleinikov, Nikolay 205, 419

P
Pala, Allansius E. 429
Pankratova, Olga 442
Parshenkova, Yulia 452
Paul, Antony Shibu 58
Pavlov, Andrey 215
Peleshenko, Tatiana 461, 471, 484, 494
Pereira, Alicia Theresa 509
Pereira, Sharlene Anna 103, 509
Poddubnaya, Natalya 274
Pominova, Karolina 517
Prakash, M 632
Prakasha, G S 92, 274, 561
Prathap, Boppuru Rudra 527, 545

Q
Qadeer, Mohammed Abdul 32, 388, 397

R
Rawat, Sanskriti 561
Raza, Zahid 494
Reddy, Vandana 527, 545
Redkokosh, Kirill 251
Romancheva, Nina 576, 590
Rudenko, Andrei 604
Rudenko, Marina 123, 131, 205, 604
Rumyantsev, Ivan 576
Rusanov, Mikhail 215

S
Sajid, Mohammad 182, 461
Savchenko, Dmitry 310
Shahid, Mohammad 1, 310, 661
Shushara, Tatiana 337
Singh, Priya 661
Singla, M. K. 748

Singla, Manish 735
Sinitsyn, Anton 136, 149
Smirnova, Daria 673
Suleimenov, Ibragim 112
Sumbwanyambe, Mbuyu 289
Suresh Manic, K 646
Swetha, V 708

T
Timofeev, Ruslan 205
Tokmakova, Marina 44, 58, 632
Trifonov, Victor 728
Trofimov, Danil 673

U
Usanov, Alexander 685
Utenkova, Maria 696

V
Vartanova, Liana 350
Velmurugan, S 375, 632
Vinodha, D 83, 103, 196, 429, 685
Vitulyova, Yelizaveta 112

X
Xamidov, M. 748
Xamidov, Munis 735

Y
Yarushev, Sergey 646
Yezhova, Elena 92
Yurkevich, Evgeniy 590

Z
Zapriy, Maria 673
Zavyalov, Ilya 718
Zayeem, Mohammad 182
Zubkov, Alexander 545
Zuev, Sergei 728