VOLUME 15 NUMBER 2 BANJA LUKA, DECEMBER 2025 (81-176) ISSN 2232-9625 (Print) ISSN 2233-0194 (Online) UDC 004

Journal of Information Technology and Applications

(BANJA LUKA)

JITA

Exchange of Information and Knowledge in Research

APEIRON

THE AIM AND SCOPE

The aim and scope of the Journal of Information Technology and Applications (JITA) is:

- to provide international dissemination of contributions in field of Information Technology,
- to promote exchange of information and knowledge in research work and
- to explore the new developments and inventions related to the use of Information Technology towards the structuring of an Information Society.

JITA provides a medium for exchanging research results and achievements accomplished by the scientific community from academia and industry.

By the decision of the Ministry of Education and Culture of the Republic of Srpska, no.: 07.030-053-160-4/10 from 3/3/2010, the journal „Journal of Information Technology and Applications“ Banja Luka is registered in the Registry of public organs under the number 591.
Printed by Markos, Banja Luka in 300 copies two times a year.

Indexed in: LICENSE AGREEMENT, 3.22.12. **EBSCO** Publishing Inc., Current Abstracts

ebscohost.com
erihplus.nsd.no
scholar.google.com
doisrpska.nub.rs
crossref.org
road.issn.org
citefactor.org
cosmosimpactfactor.com

Printed on acid-free paper

Annual subscription is 30 EUR
Full-text available free of charge at http://www.jita-au.com

UDC 004

ISSN 2232-9625 (Print)
ISSN 2233-0194 (Online)

JITA

VOLUME 15 NUMBER 2 BANJA LUKA, DECEMBER 2025 (81-176)

Contents

Dear Readers,

With great pleasure, we present to you the second issue of Volume 15 of the Journal of Information Technology and Applications (JITA)—an issue that marks not only the continuation of our publishing cycle but also the steady growth of our scientific community. This issue reflects the diversity and dynamism of contemporary ICT developments and applications. The published papers span a wide range of topics—from IoT architectures and embedded systems to artificial intelligence, machine learning, digital education, and applied data analytics. Such thematic breadth is a testament to the interdisciplinarity of ICT and its growing impact on all segments of society.

The research presented in this issue includes:
- innovative IoT gateway architectures and decentralized data collection models;
- advances in CNN-based intelligent transport systems and driver-assistance technologies;
- applications of low-cost sensors and AI in environmental monitoring;
- the role of AI assistants in contemporary pedagogical practice;
- case studies of AI-driven business transformation;
- development of domestic language AI models and insights into the evolution of large language models;
- analyses of financial sustainability of e-learning platforms;
- improvements in distance learning systems.

These papers show the rapid integration of intelligent systems into daily life and highlight the importance of ICT research in shaping future socio-technical systems. In this issue, young researchers and early-career academics stand out, demonstrating technical skill and creative problem-solving. We actively support such authors as a core mission of our journal.

As JITA expands its international editorial board, indexing, and visibility, we reaffirm our commitment to open scientific exchange, methodological transparency, and the spread of research that advances theory and practice. We invite researchers from academia and industry to contribute, join projects, and strengthen our community.

We thank all authors, reviewers, and members of the editorial and technical teams for their dedicated work. Their contributions ensure the high quality and integrity of JITA and enable the journal to remain a respected forum for ICT research.

We wish you an inspiring reading experience and look forward to receiving your future submissions.

Editor-in-Chief
Dalibor P. Drljača, PhD
Pan-European University APEIRON
Journal of Information Technology and Applications (JITA)

Implementation of a Meshtastic Gateway System With a Local Database for Iot Applications

Daniel Menićanin[1], Jelena Radanović[2], Dražen Marinković[3]

[1]*Pan-European University Apeiron, Faculty of Information Technology, Banja Luka, Bosnia and Herzegovina, danijel.menicanin@gmail.com, 0009-0001-6311-4043*

[2]*Pan-European University Apeiron, Faculty of Information Technology, Banja Luka, Bosnia and Herzegovina, dev.radanovic@gmail.com, 0009-0001-5135-7662*

[3]*Pan-European University Apeiron, Faculty of Information Technology, Banja Luka, Bosnia and Herzegovina drazen.m.marinkovic@apeiron-edu.eu, 0009-0006-8001-2168*

Original scientific paper

https://doi.org/10.7251/JIT2502085M UDC: 004.76+004.42

Abstract: This paper presents a system for reliable collection, filtering, and processing of data from a LoRa Meshtastic decentralized network, developed for use in remote areas with weak or no mobile network coverage. The core idea stems from the need to enable efficient exchange of small data packets at intervals, without relying on expensive and often unavailable internet infrastructure. The key innovation lies in the implementation of the Meshtastic gateway concept, which provides internet access via HTTP requests, while the developed database model ensures continuity and reliability of data transmission. Data arriving in the network as unstructured messages are extracted using regular expressions, transformed into JSON format, and sent to the visualization platform Grafana, while simultaneously being stored in a local database for later queries, research, and analysis. The system's reliability is further enhanced by introducing a two-layer acknowledgment mechanism (Meshtastic ACK_APP and remote API application-level ACK), as well as an offline mode that logs undelivered messages and their causes through flags and error records. This ensures resilience to data loss and enables seamless operation continuation after connection interruptions.

Keywords: LoRa, Meshtastic, IoT, Database

Introduction

The Internet of Things (IoT) has in recent years become a key technology in areas such as industrial automation, smart cities, agriculture, and environmental monitoring. The fundamental idea of IoT systems is the collection and processing of data through a network of sensors and their transmission to remote platforms where analysis and visualization take place.

However, in practice, a major problem arises, most of these systems depend on a stable and continuous internet connection, which is not always achievable in remote or infrastructure-limited areas.

The challenge is particularly evident in applications that require only periodic transmission of small amounts of data. In such cases, maintaining a permanent mobile or fixed internet connection represents an unnecessary cost and reduces the overall cost-effectiveness of the system. This creates a gap between real-world needs and existing commercial solutions. As a response to this challenge, a system was developed that combines the LoRa [1] Meshtastic network and a multiplatform application for data collection and processing. Unlike LoRaWAN solutions, which require a centralized gateway and a connection to a provider's network, Meshtastic uses a mesh topology. This means that each node in the network functions both as an endpoint and as a repeater, achieving resilience to interruptions and eliminating dependence on a single central point. In this way, the system can operate completely offline, without the need for infrastructure support.

The role of the developed application is to receive data from the Meshtastic network, perform filtering, and transform it into a structured JSON format. Since a large number of non-system and auxiliary

messages appear in the network, regular expressions (regex) are used to extract only relevant data from predefined senders. The filtered data is then written to a local database and simultaneously sent to the Grafana [2] visualization platform via HTTP API calls. This ensures dual security, data is available in real time and permanently stored locally for further analysis. Figure 1 shows the Grafana interface used for data visualization.

Figure 1. *Visualization of collected data using the Grafana dashboard*

The key contribution of the system lies in the implementation of reliability and redundancy mechanisms. The Meshtastic network provides delivery confirmation within the protocol itself (ACK_APP), while the application further utilizes HTTP responses from remote APIs. In case of internet disconnection, all data is automatically stored in the local database with a clearly marked status "undelivered" and an associated error reason. When the connection is restored, the system automatically performs synchronization and retransmission, eliminating the possibility of data loss.

A special quality of the developed solution lies in the fact that the application is cross-platform, developed in the Dart/Flutter environment, and runs on Windows, Linux, and Android operating systems. This provides high flexibility of use from server operation in office environments, through field monitoring on laptops, to mobile deployment on Android devices.

Based on the problem identified and the proposed solution, this work seeks to answer the following research questions:

1. How can IoT systems ensure reliable and cost-effective data transmission in environments with limited or no internet infrastructure?
2. To what extent can a decentralized LoRa-based mesh network (Meshtastic) serve as a viable alternative to traditional LoRaWAN or cellular solutions?
3. How effective is regex-based filtering in extracting relevant sensor data from a noisy, unstructured message stream?
4. What level of reliability and redundancy can be achieved through hybrid local and remote data handling, especially during intermittent connectivity?
5. How does a cross-platform gateway application contribute to the practical deployment and scalability of offline-capable IoT systems?

METHODS AND MATERIALS

The architecture of the developed system is designed to integrate all the key layers required for reliable data collection, processing, and distribution in environments where internet infrastructure is unavailable or not economically viable.

At the base level of the system are sensors connected to ESP32 microcontrollers. These devices are responsible for collecting measurement values, formatting them into key–value pairs, and preparing them for transmission. The messages then enter the LoRa Meshtastic network, which represents the communication layer of the system. Thanks to its mesh topology, the network is self-sustaining and allows messages to be dynamically routed through multiple nodes until they reach their destination.

When messages reach the node connected to the application, the data processing layer begins. The application retrieves all incoming messages from the network via serial communication and filters them using regular expressions. After filtering, the data is transformed into JSON format, making it standardized and ready for integration with remote services or storage in the local database.

In the third layer, data distribution and storage take place. The data is sent to the remote Grafana service via HTTP API calls, while simultaneously being written into the local database. The database does not serve merely as passive storage, it actively participates in system functionality by enabling analyti-

cal queries and reconstruction of historical events. A special feature is the flag-based record marking mechanism: if data is not successfully delivered to the API, it remains in the database marked as "undelivered" with the corresponding error reason recorded, ensuring complete transparency. Figure 2 shows the data flow diagram of the system architecture.

Figure 2. *LoRa Meshtastic Gateway System – Data Flow Diagram and Acknowledgment Structure*

One of the key aspects of the system architecture is the implementation of a two-layer message delivery confirmation mechanism.

At the protocol level, the Meshtastic network ensures message receipt confirmation through the ACK_APP port number. On top of this, an additional application layer is implemented, where the remote API server returns an HTTP response indicating the success of data processing. This combination allows the system to maintain a clear record at all times of whether a message has been successfully delivered and acknowledged.

If an internet connection interruption occurs, the application continues to operate smoothly in offline mode, during which all data is stored in the local database and marked with an error flag. Once the connection is restored, the application automatically retries transmission and synchronizes the database with the remote service.

Furthermore, the system architecture has been extended to include two operational modes of the application: server mode and client mode.

Figure 3 shows the application in server mode.

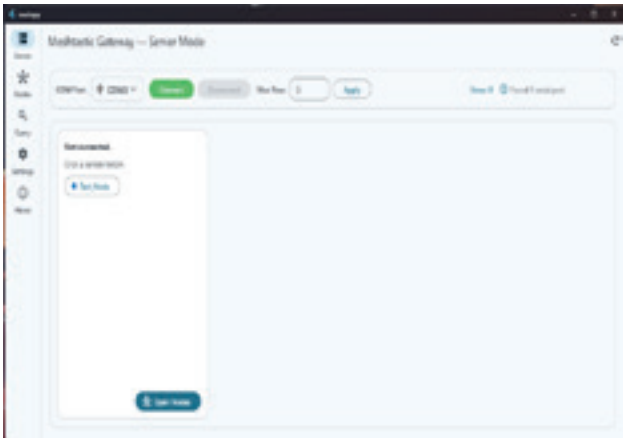


Figure 3. *Meshtastic Gateway - Server mode*

In server mode, the application assumes a central role. It serves as the point where complete data processing is performed, including filtering, transformation into JSON format, database entry, and transmission to Grafana. This mode of operation implies that the server application maintains the main instance of the database and functions as a gateway between the Meshtastic network and the internet [3]. In this way, the server mode enables integration of the entire system with external analytical tools and ensures centralized real-time monitoring of all network nodes.

In client mode, the application operates in a simplified configuration. It does not process or store data directly but instead sends queries through the Meshtastic network to the main application running in server mode and retrieves results from the database. This allows remote users, those without direct internet access, to obtain information about the system's status. The client mode extends the concept of decentralization by enabling access to data even in situations where the client does not have a connection with the server through a traditional network, but exclusively via the Meshtastic network. This functionality is particularly important for field operations, where operators or researchers can monitor

data on their computers or mobile devices, while the main server remains located in a secure environment with constant power supply and internet access.

The visual and user interface component of the application has been carefully developed to ensure intuitive operation in both modes. The Dashboard provides an overview of all active nodes and key system metrics, while the configuration section allows users to enter all parameters necessary for operation (node IDs, names, API keys, and color categorization). A significant feature is the integration with the database through the application interface, which enables users to explore and query data locally without connecting to remote services. Application settings, such as theme selection or automatic connection to the first available node upon startup, further simplify usage in field conditions.

By combining server and client modes, the system architecture provides complete flexibility and achieves a balance between centralized processing and decentralized access. In this way, the developed solution gains characteristics that surpass traditional IoT implementations: the system becomes simultaneously robust, scalable, and adaptable to diverse operational conditions.

The developed application is designed to give users full control over system operations and access to all collected data, with special attention paid to interface clarity and usability. Its functionalities are organized into several modules that together form a single, intuitive tool for managing distributed IoT nodes.

At the core of the application is the Dashboard, which serves as the system's operational center. On this screen, users can view all active nodes, their basic parameters, and statuses. For each node, data such as identification number, name, and the latest received data set are displayed, along with availability indicators. Special attention has been devoted to presenting the overall "health" of the system. Users can monitor node availability, message loss rates, transmission delays, and general communication status. The Dashboard also includes the status of the local database and the remote API, allowing users to see in real time whether data is successfully transmitted to the analytical platform or held in a waiting state due to internet issues.

Visual alerts notify users of critical events such as network interruptions, API connection loss, or sudden drops in node battery voltage and capacity. This enables quick assessment of the entire system and timely response to any problems.

The second segment of the application is the Node Configuration Module, where all parameters necessary for data monitoring and processing are defined. For each node, it is possible to enter its SENDER ID, a name for easier identification, and an API key [4] that enables communication with remote services. In addition, Node ID values are assigned, which are used on the visualization platform side, along with color labels that categorize nodes within the interface to improve clarity in more complex systems.

This section provides full control over the integration of new nodes, while the flexibility of the module allows for easy addition, modification, or removal of devices from the network.

Figure 4 shows the configuration interface for the nodes.

Figure 4. *Configuration interface for the nodes*

The third functional unit of the application relates to database operations, implemented through the Query Module. Here, users can explore and analyze historical data stored in the local database. The data can be filtered by time intervals, sender identity, or type of measurement. This functionality enables users to generate reports and perform analyses even when the internet is unavailable or when remote services are not operational. In this way, the database is not merely a passive element of the system but becomes an active tool for exploring and evaluating all collected values.

The Settings Module forms the fourth functional unit and is designed to adapt the application's opera-

tion to user needs and specific conditions. It allows for theme selection to improve visibility under various lighting environments and the activation of an auto-connect option that links to the first available node on the serial port at startup, significantly simplifying field operations. This section also provides fine control over security settings, including the acceptance of self-signed SSL certificates when working with private or test API services.

Another important part of the application is the About section, which provides a summary of essential information about the application and its authors. This section serves as an overview of the basic documentation and metadata, giving users insight into the software version and providing contact information for technical support.

All system functions are integrated into a single application that combines monitoring, configuration, and data analysis. This unified design provides users with a practical and reliable tool for managing and supervising distributed IoT systems.

In distributed IoT systems that rely on LoRa and Meshtastic networks, message loss and communication interruptions represent real challenges. Therefore, the developed solution places particular emphasis on reliability and redundancy. Mechanisms are implemented across multiple layers, from the transport layer managed by Meshtastic to the application layer and local data storage.

Figure 5 shows the Meshtastic logger, which reads the data transmitted by a node via the COM port.

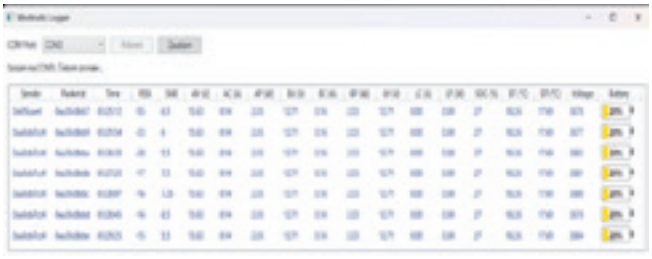


Figure 5. *Meshtastic logger*

The ACK [5] mechanism within the Meshtastic network represents the first layer of reliability. Each transmitted message can receive an acknowledgment (ACK) from the next node, ensuring that the data has been successfully forwarded within the network, even if it has not yet reached the final application. This level of confirmation is particularly important in mesh topology, where messages may be routed through multiple intermediate nodes. If an ACK is not received, the message is retransmitted, thereby reducing the likelihood of data loss due to transient interference.

Building on this, an application-level ACK is implemented, originating from the remote API after an HTTP request. While Meshtastic confirms only that the message has successfully traversed the network, the HTTP response provides information about whether the data was actually received and recorded by the external service, such as Grafana, as shown in Figure 6.

Figure 6. *Grafana*

If the remote server returns an error or fails to respond within the expected time, the application logs this event and marks the message as “undelivered.”

To ensure data redundancy, every message, whether successfully sent to the API or not is recorded in the local database. If external transmission fails, the message is flagged as “undelivered,” and the reason for failure (unavailable internet connection, timeout, or API response error) is stored. This provides a complete record and enables straightforward debugging and post-analysis. [6]

When the internet connection is restored, automatic synchronization is triggered. The application identifies all messages marked as “undelivered” and resends them sequentially to the remote API until a positive ACK is received. This mechanism effectively eliminates the risk of permanent data loss and allows the system to continue operating seamlessly even after prolonged network interruptions.

Such a multi-layer reliability approach makes the solution suitable for real-world scenarios where com-

munication disruptions are frequent, such as remote rural areas or industrial environments. [7]

RESULTS

The developed system was implemented through a combination of hardware and software components, where ESP32 microcontrollers with attached sensors formed the foundation for data collection, while the LoRa Meshtastic network provided the communication layer. The software part included an application developed in the Dart/Flutter environment, a parser for data filtering and transformation, and a local database in which all messages were stored along with their delivery statuses. This created a self-sustaining mechanism capable of operating under various network conditions.

In the first phase of testing, the system operated under stable conditions with a constant internet connection. Sensor nodes sent data at 10-second intervals, and the application processed and forwarded them to the remote API. The latency from the moment of data generation on the sensor to its display in Grafana averaged 1.2 seconds. During several hours of continuous operation, no data loss was recorded, confirming that the system functions flawlessly in ideal network conditions.

Next, the system was tested in scenarios involving internet connection interruptions. During periods lasting between five and thirty minutes without connectivity, all data were smoothly recorded in the local database, marked as "undelivered," and supplemented with metadata about the cause of the error. Once the connection was restored, the application automatically performed resynchronization and forwarded all messages to the remote API. The results showed that no data were lost, and the local database allowed complete event reconstruction and precise tracking of the time each error occurred.

The network's performance was also tested with a ten nodes. Nodes were included, periodically sending data to simulate a more complex system. The application's parser successfully filtered out around 35% of the traffic generated by Meshtastic in the form of auxiliary and non-system messages, so only relevant data from defined senders proceeded for processing. Despite the increased communication volume, the application maintained stable operation, with no network congestion or message loss observed.

The local database proved to be not only essential for maintaining system integrity but also invaluable for data analysis. Queries such as extracting all data for a specific node within the last 24 hours or searching for voltage values above a defined threshold were executed in under 100 milliseconds. This confirmed that the database serves not merely as a passive layer but as an active tool for real-time exploration of historical data.

The user interface demonstrated clarity and usability during practical testing. The Dashboard provided a clear overview of active nodes and key system metrics, while the Configuration Module offered flexibility in assigning API keys, identification numbers, and visual markers.

The Query Module proved especially valuable in offline mode, enabling report generation directly from the local database without the need for internet access. Features such as automatic connection to the first available node upon startup were particularly useful in field conditions, where speed and simplicity are crucial.

The research results confirm that the system provides stable and reliable performance. Under stable network conditions, the average latency was 1.2 seconds, no data loss occurred in any scenario, and around 35% of non-system traffic was effectively filtered out. The network successfully handled the load of ten nodes, and the local database enabled fast analytical queries with response times below 100 milliseconds. The main limitation remains the low bitrate of LoRa communication, making the system optimal for small amounts of data transmitted at periodic intervals.

DISCUSSION

The evaluation of the implemented system indicates that the integration of a LoRa Meshtastic mesh network with a locally hosted data processing and storage application constitutes a viable and reliable solution for Internet of Things (IoT) deployments in environments with limited or intermittent internet connectivity. The system architecture, based on decentralized communication and dual-layer reliability mechanisms, demonstrated robustness in various test scenarios, including network interruptions and increased communication load.

A significant contribution of the proposed solution lies in its ability to operate autonomously in

offline conditions. The introduction of a two-layer acknowledgment mechanism, protocol-level ACKs via the Meshtastic network and application-level acknowledgments from HTTP API responses, ensured data integrity across all stages of transmission. In cases of connection loss, data packets were securely stored in a local database, appropriately flagged as undelivered, and subsequently synchronized upon reconnection. This redundancy model significantly reduces the risk of permanent data loss and supports system resilience in real-world conditions.

In comparison to conventional IoT architectures, particularly those based on LoRaWAN, the presented solution eliminates the dependency on centralized gateways and service provider infrastructure. The adoption of a mesh topology ensures fault tolerance through dynamic rerouting and enables flexible deployment in areas with challenging topography or limited access to power and network infrastructure.[8]

Furthermore, the implementation of server and client operational modes enhances system scalability and decentralization, allowing field users to access critical data via the Meshtastic network without direct access to the internet.

Although effective in many cases, the LoRa technology still has several built-in constraints. The restricted data rate and narrowband communication capacity limit the applicability of the system to use cases involving low-volume and periodic data transmission. While this makes it suitable for scenarios such as environmental monitoring, agriculture, or remote infrastructure supervision, it is not applicable in contexts requiring high-throughput data exchange or multimedia support.

Another notable constraint pertains to message parsing and data filtering. Although the system effectively utilizes regular expressions to eliminate non-system and auxiliary traffic, future iterations may benefit from the implementation of more adaptive parsing methods, such as machine learning-based classification or context-aware filtering, particularly in larger and more heterogeneous networks.[9]

Potential directions for future development include:

1. Optimization of performance in high-density node environments,
2. Support for alternative or parallel visualization platforms,
3. Implementation of real-time network topology mapping,
4. Enhancements in data security, including encryption and authentication layers,
5. Remote management and over-the-air firmware updates for sensor nodes.

Overall, the system demonstrates a high level of operational stability, adaptability, and practical value in field deployments. The modular software architecture, real-time visualization integration, and effective handling of intermittent connectivity represent key advancements in the design of resilient and cost-efficient IoT systems.

These findings suggest that the proposed approach provides a sustainable foundation for the further development of decentralized sensor networks tailored to infrastructure-limited environments.[10]

Conclusion

The developed system demonstrated that the combination of a LoRa Meshtastic network and an application for filtering, transforming, and distributing data can provide a reliable and cost-effective infrastructure for IoT applications in areas with weak or no mobile network coverage. Thanks to the mesh topology and the ability of each node to function simultaneously as both a transmitter and a repeater, the network ensured resilience to interruptions and stable message transmission.

Meanwhile, the application contributed by eliminating non-system data, standardizing the content into JSON format, and enabling integration with remote services such as Grafana.

Testing confirmed that the system operates successfully even in conditions with internet connection interruptions, as the local database assumes the role of maintaining system integrity. The delivery confirmation mechanisms at both the network and application levels guaranteed that no data were lost, while later resynchronization ensured consistency between the local database and the remote API. A key advantage of the solution lies in its flexibility, the same system can be used in server mode, as a central point for processing and visualization, or in client mode, where remote users can access data through Meshtastic even without an internet connection.

The system's limitations are primarily related to the capacity of LoRa technology, which is suitable for

small amounts of data transmitted periodically but not intended for applications requiring high throughput. Nevertheless, in scenarios involving intermittent and low-volume transmissions, such as environmental monitoring, agriculture, remote industrial site supervision, or smart city applications in poorly connected areas, the developed solution provides an optimal balance between reliability, simplicity, and cost-effectiveness.

REFERENCES

[1] F. Freitag, J. M. Solé, and R. Meseguer, "Position Paper: LoRa Mesh Networks for Enabling Distributed Intelligence on Tiny IoT Nodes," 2023. [Online]. Available: https://www.researchgate.net/publication/371826243

[2] G. Y. Kusuma and U. Y. Oktiawati, "Application Performance Monitoring System Design Using OpenTelemetry and Grafana Stack," Journal of Internet and Software Engineering, vol. 3, no. 1, 2024.

[3] R. P. Centelles, R. Meseguer, and F. Freitag, "Exploring Open Source and Proprietary LoRa Mesh Technologies: A Minimalistic Routing Protocol for LoRa Mesh Networks," 2024. [Online]. Available: https://www.researchgate.net/publication/380253012

[4] R. Berto, P. Napoletano, and M. Savi, "A LoRa-Based Mesh Network for Peer-to-Peer Long-Range Communication," 2021. [Online]. Available: https://www.researchgate.net/publication/352707672

[5] M. A. Khan, M. T. Islam, and A. R. Chowdhury, "Implementation of Multi-Hop Mesh Networking Using ESP32 for IoT Communication," 2024. [Online]. Available: https://www.researchgate.net/publication/389763039

[6] T. G. Durand, "Performance Evaluation of a Mesh-Topology LoRa Network," *Sensors*, vol. 25, no. 5, 2025.

[7] D. Arregui Almeida et al., "Gateway-Free LoRa Mesh on ESP32: Design, Self-Healing and Multi-Hop Communication," *Sensors*, vol. 25, no. 19, 2025.

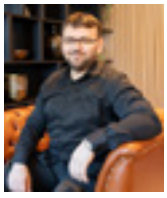
[8] P. Kietzmann, J. Alamos, D. Kutscher, T. C. Schmidt & M. Wählisch, "Long-Range ICN for the IoT: Exploring a LoRa System Design," *Proc. 21st IFIP Networking Conference*, 2022.

[9] N. L. Giménez et al., "Embedded federated learning over a LoRa mesh network," *Elsevier*, 2023.

[10] J. R. Cotrim, J. H. Kleinschmidt & al., "LoRaWAN Mesh Networks: A Review and Classification of Multi-hop Communication," *Sensors*, 2020

Received: November 3, 2025
Accepted: November 18, 2025

ABOUT THE AUTHORS



Daniel Menićanin, a student at the Faculty of Information Technology specializing in Programming and Software Engineering at Paneuropean University Apeiron, is an innovator recognized for developing groundbreaking solutions. Among his notable projects is the RC Platform Controlled via PlayStation 4 Controller, designed to support children with developmental challenges. His innovation, Software Development for CNC Hydraulic Presses, earned him a gold medal at the INN&TECH conference in Sarajevo, Bosnia and Herzegovina. In recognition of his exceptional contributions to education, science, and the arts, Daniel received the Plaque of the City of Banja Luka. He was also awarded the prestigious Butterfly Innovation Award in the Youth category, presented by the Regional Cooperation Council. Currently, Daniel's research focuses on robotics, industrial machinery, and automation, further advancing his expertise in developing cutting-edge technological solutions.

Jelena Radanović, a dedicated and accomplished student of Programming and Software Engineering at Paneuropean University Apeiron, is widely recognized for her innovative contributions and exceptional achievements in the field. She has received the prestigious Butterfly Innovation Award and the INOST Medal for her project Press Brake Software, showcasing her expertise in software development and creative problem-solving.
Her work demonstrates a unique blend of technical proficiency and creativity, solidifying her reputation as a rising talent in programming and software engineering. Jelena continues to push boundaries in her studies and projects, positioning herself as a leader in technological advancement and innovation.

Dražen Marinković was born in 1978. He received his M.Sc. degree in 2015 and Ph.D. degree in 2020, both from the Faculty of Informatics at the Pan-European University Apeiron in Banja Luka, specializing in computer and informatics engineering. He is currently an associate professor at the Faculty of Informatics, Pan-European University Apeiron. His research interests include data science, computer networks, and related fields in modern computing technologies.

FOR CITATION

Daniel Menićanin, Jelena Radanović, Dražen Marinković, Implementation of a Meshtastic Gateway System With a Local Database for Iot Applications, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:85-92, (UDC: 004.76+004.42), (DOI: 10.7251/JIT2502085M), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

CNN-Based Road Sign Recognition for Driver Assistance

Boris Borovčanin[1], Samed Jukić[2]

[1] *Department of Information Technologies, Faculty of Engineering, Natural and Medical Sciences, International Burch University, Sarajevo, Bosnia and Herzegovina, boris.borovcanin@stu.ibu.edu.ba, 0009-0002-7993-0544*

[2] *Department of Information Technologies, Faculty of Engineering, Natural and Medical Sciences, International Burch University, Sarajevo, Bosnia and Herzegovina, samed.jukic@ibu.edu.ba, 0000-0001-7931-4093*

Original scientific paper

https://doi.org/10.7251/JIT2502093B UDC: 629.3.067:004.032.26

Abstract: Considering established relevance to the GTSRB dataset, it is important to emphasize that research investigates the effectiveness of convolutional neural networks (CNN) in the field of road sign recognition. Following that wide range of techniques for comprehensive preprocessing pipelines were implemented, including data normalization and augmentation as well as resizing images. The CNN model has demonstrated the ability to overcome adverse conditions across multiple road sign classes, demonstrating outstanding scores against the performance metrics used in testing and evaluation process. Model achieved classification accuracies exceeding 99% across most categories. Nevertheless, in certain classes there is presence of performance metric decline related to the inaccurate visualization and contradiction of features. The crucial role of the preprocessing phase has been highlighted while the implementation of the CNN model has been identified as one of the most reliable approaches in the field of road sign recognition. However future implications must be considered to achieve the full potential of the model. Some of the crucial contributions for the future will be introducing real life variation in the dataset. On the other hand, occlusion, lighting and weather conditions are the important factors that should be brought into focus.

Keywords: road sign recognition, machine learning, convolutional neural networks, adas

Introduction

One of the critical components of Intelligent Transportation Systems is Road Sign Recognition, whose purpose is to provide support for digital components of the vehicle in the field of traffic regulations interpretation. Taking it into account the examples of these interpretations are speed limits, hazard warnings, and navigation commands. Following that Road Sign Recognition for Driver Assistance takes into consideration important aspects such as efficiency and management of traffic flow, adherence to regulations and road safety. The real-time recognition of road signs in an Advanced Driver-Assistance System (ADAS) [1] reinforces the process of human error suppression such as missing important signals or misinterpreting them. However, the system enhances the level of situational awareness of the drivers while allowing them to make effective decisions in order to minimize chances of potential accidents.

Road sign recognition forms the difficulty of perception systems in autonomous vehicles, enabling safe navigation in complex environments. Autonomous vehicles comply with the laws by monitoring traffic signs and consequently to that assist in coordinating the interactions with pedestrians and human-driven vehicles. Public safety issues in this concept are the subject of address, while enhancing the more extensive adoption of autonomous driving technologies all over the world. It is important to emphasize that there is evident utilization of Convolutional Neural Networks (CNNs) [2] in this project, given that these networks are specialized in extracting spatial data, consequently making them ideal for classification problems involving images.

Attention mechanisms are one of the advanced techniques that improve the model's capacity to focus on important sign regions while at the same time not compromising performance of computation. Taking that into account it is obvious that all these necessary approaches are implemented using on-the-fench

algorithms that process data in real-time, with the main purpose for deployment in dynamic real-world environments. By realizing complicated and diverse conditions in the real world, the system differentiates itself. Regarding that, the process includes reliable operating systems under all circumstances like variations of light, bad weather, and the presence of occlusions. Considering these issues, solutions are essential for developing a framework that can move applications across different environments.

This research contributes to the field by evaluating CNN-based road sign recognition using the GTSRB dataset, with emphasis on preprocessing strategies and performance metrics. The study aligns with global trends in sustainable urban mobility and intelligent vehicle technologies, bridging the gap between human-driven and AI-driven systems.

In the context of contribution to road safety and compliance with traffic regulations, it is important to emphasize that timely and precise availability of road sign recognition reduces the risk of driver distraction. The focus is on adherence of traffic regulations for the purpose of safe and legal driving behavior. Considering the above, the project corresponds with the global trends and contributes to achieving sustainable urban mobility and the implementation of intelligent vehicle technologies, as it improves road safety. It is important to emphasize that the integration of road sign recognition in the fully automated and semi-automated system closes the gap between the vehicle operated by a human and the one operated by artificial intelligence.

Taking it into consideration, among the greatest challenges that development of autonomous driving systems faces in the current decade is the reliable identification and real-time recognition of road signs under the impact of a wide range of different environmental circumstances, including various weather conditions. Road signs are critical indicators for navigation consistency and legal compliance with traffic rules. Misinterpretations or delay in recognition could undermine the performance and implementation of ADAS. In order to address the problem of misclassification, CNN model has been trained on GTSRB dataset for the purpose of ensuring efficient real-time road sign recognition, measured using performance metrics. As the result of data preprocessing and optimization of the pipeline, the model could decrease performance efficiency. This research provides a comprehensive structure of research in neural networks and deep learning; however, it also has a direct impact on the development of the automotive industry by improving reliability and reducing driver mistakes in autonomous systems.

The section Literature review will include reviews of 10 research papers in the context of their research purpose, methodology, and findings and its relation to the topic of proposed investigation. The proposed dataset is going to be subject of training and assessment, including the stages of preprocessing data, model training and its evaluation. In the Dataset and Methodology section the theoretical background will be explained in relation to methodology. Sections related to Results and Discussion are going to contain the analysis and comparison related to the contributions and limitations of the model. The reference is going to be to the tables and figures where collected data is going to be sorted. Afterwards the whole research paper is going to be summarized.

LITERATURE REVIEW

In the context of literature related to the topic of road sign recognition it is important to take into consideration a variety of approaches related to the topic of research. There is evident contribution of recent research to find out methodologies addressing recognition, accuracy, and reliability issues in real time. However, the following investigations are implementing techniques including deep learning models, classic approach to machine learning as well as optimization, while providing a comprehensive overview and discussion. The alignment of the following papers to corresponding research is in the field of efficiency and accuracy of machine learning models for the purpose of advancement in real time road sign recognition while developing driving safety as well as improving autonomous driver assistance systems.

It is important to emphasize that a structured protocol was employed for the purpose of identifying, selecting, and synthesizing the reviewed literature. Considering that these relevant studies were retrieved from IEEE Xplore, ScienceDirect, SpringerLink, and Google Scholar using keywords associated with road sign recognition and deep learning. Taking it into account the inclusion criteria required, the peer-reviewed research papers are interpreting pro-

posed machine-learning or deep-learning methods in the field of road sign detection or classification, reporting quantitative performance, and interpreting clearly defined datasets. Evaluating the abstracts as well as complete texts, the final research papers were selected based on relevance experimental results, methodological quality, or relevant computational modeling in the field of road sign recognition. On the other hand, work that does meet the criteria that was subject of evaluation has not been taken into consideration.

Observing the approach introduced in **[3]**, it is important to emphasize that the purpose of this research is performance evaluation while addressing issues related to real-time road sign recognition. This process is determined by implementation of CNNs. Through the analysis of this research, it has been discovered that methods of this investigation are based on development of transfer learning, which was a crucial element related to GTSRB evaluation presented in [4]. The results have shown that the accuracy rate of CNN reaches 98.9%, suggesting substantially reduced computational cost. When it comes to the relevance of the approach [4] to proposed research, it could be scientific proof of efficiency related to the CNNs in the field of road sign recognition, while reinforcing the theory of accurate and safe models development.

When it comes to the work discussed in [5] the main focus is on the advancement of YOLO (You Only Look Once) models which are related specifically for traffic sign recognition in complex environments, in fact circumstances which are not controlled. Taking it into consideration, the YOLOv4 modified model emphasizes pyramid networks method as an important feature while enhancing effective detection under low light conditions. The outcome has shown advancement in precision and recall results, while accuracy rate reached 95% under more challenging occasions. This could be useful for proposed research in the field related to ensuring reliability associated with real-time sign recognition for driver assistance, while optimizing YOLO for complex environments.

The focus of research discussed in [6] the research paper is on the multimodal data creation and implementation related to the traffic signs recognition under the impact of comparative weather conditions. They have utilized a combination of CNNs and RGB images as main methods, while considering thermal data to perform data fusion. The results have shown improvements in recognition accuracy that have risen by 7%, in comparison with RGB methods alone where the environment was dark and cloudy. The relevance of the research to the implementation of corresponding work is discussed in [6] in the context of combining infrared with RGB data in order to improve recognition under the impact of poor lighting, while focusing on achieving safe driving conditions.

Key idea for idea introduced in [7] is establishing attention models to enhance more efficient optimization and acquisition around traffic sign identification models. In order to achieve that idea there has been an elevated technique of self-attention in compliance with CNN architecture. At the end the outcome suggested a rise in accuracy of detection in the field of partially covered and small signs, with a high percentage of F1 score equal to 97.8%. Considering the purpose of proposed investigation and the idea of the work discussed in [7], the alignment is emphasized in the domain of elevating reliability, while implementing attention mechanisms in order to enhance detection of small or relatively covered signs.

When it comes to the idea introduced in [8], the aim was to alter MobileNet compact architecture related specifically to the low-power devices. There have been techniques related to aggressive data augmentation which include training of MobileNet models on the dataset presented in [4]. The findings have shown that accuracy rate achieved 95.5% under the circumstances where memory consumption has been reduced in order to be suitable for embedded systems. It aligns with the focus of proposed research on implementation of real-time applications which are resource-efficient, suitable for autonomous vehicles as well as the Advanced Driver-Assistance System (ADAS).

For the purpose of examination of work presented in [9], regarding the methodological approach in this specific research, there has been evident comparison of the traditionally based approach such as SVMs and K-NN against the deep learning models used to interpret dataset. The traditional method is mainly based on predefined elements, since relevant features are manually extracted. The outcome has shown CNNs as the most relevant approach to this kind of data, since the MLP model reached an accuracy rate equal to

98.98%, while the CNN achieved accuracy of 99.46%, on the specific dataset. The relevance of this specific research to the topic of corresponding research is shown in the concept that includes evaluation of traffic sign recognition systems in order to support development of ADAS.

Keeping track of demonstrating the efficacy of the deep neural networks was in the focus of work introduced in [10] in order to implement reliable multi-class traffic sign recognition. In the context of methods used it is important to emphasize that this approach interprets deep neural networks in the combination with application to the dataset as reported in [4]. The results have shown that accuracy rate reached 99.46%, emphasizing differences to the previously used methodologies. It is important to take into account that the approach of research discussed in [10] aligns with proposed research in the field of implementation of deep neural networks in order to achieve high accuracy rate, emphasizing application of deep learning in real time in order to achieve road safety.

Research presented in [11] aims to explore the implementation of traffic sign recognition, while elevating the hierarchical classification method. Following that there has been utilized a combination of hierarchical classification and cascade classifier techniques. Considering the results it is important to take into account that classification accuracy reached a rate equal to 98.96% on the GTSRB dataset reported in [4]. This research aligns with the purpose of authors' work in the context of improving accuracy rates in the real time, complex traffic environments.

Considering the work presented in [12] the focus was on evaluating model performance of traffic sign recognition under the impact of data augmentation. Methods that acknowledge this approach include applied transformation techniques such as scaling, rotation and adaptation of brightness, implemented to train data. The outcome suggests accuracy rate that achieves 98.9%, while highlighting robustness of the performed techniques. This research is related to proposed investigation within the framework of robustness of potential implementation in real time driving conditions.

The main idea of work proposed in [13] is the examination of the local binary patterns (LBP) application in combination with SVMs for the purpose of effective traffic sign recognition. Considering the focus of the research, there have been implemented approaches that include implementation of the LBP and SVM related techniques in order to gather handcrafted features. However, in the field of this research, it is important to emphasize that there are significant outcomes related to the accuracy rate, which has been reported as a percentage of 98.78% on the GTSRB dataset is presented in [4]. Considering the approach of research discussed in [13], there is the prospect of applying the relevant techniques in the context of this investigation.

Considering the previously reviewed literature and the resources authors have used, the research hypothesis will be stated as follows: "The Convolutional Neural Networks (machine learning) model can efficiently recognize road signs in real-time, for the purpose of autonomous and safe driving development." Such a methodology is most appropriate, since the experimental design controls the ability of the model to recognize signs effectively in real time conditions.

METHODS AND MATERIALS

Taking into consideration the dataset that is planned to be used in this project, GTSRB (German Traffic Sign Recognition Benchmark Dataset) has been chosen, as noted in [4]. Dataset originates from the 2011 ISI GTSRB competition and takes into consideration road sign images from the real-world conditions from German roads. This dataset includes 50000 images, traffic signs and its representation which has been organized in 43 categories. It is important to emphasize for the purpose of machine learning processing, images are conformed and resized to a uniform dimension which is equal to 32×32 pixels ratio.

Corresponding dataset requires ordinary preprocessing steps such as implementing fixed input size, conducting pixel values normalization while at the same time applying data augmentation. GTSRB represents robust benchmark that can be used for the purpose of evaluation and comparison, since it adopts wide range of different images. Considering the structure and characteristics, images have different height, angle and illumination, while providing a substantial dataset for testing and training.

Table 1 *Columns of GTSRB dataset: The table contains relevant information regarding metadata of images included in the dataset.*

Column	Description
Filename	Name of file
Width	Width of image
Height	Height of image
Roi.X1	X-coordinate of the top-left corner of the Region of Interest (ROI).
Roi.Y1	Y-coordinate of the top-left corner of the ROI.
Roi.X2	X-coordinate of the bottom-right corner of the ROI.
Roi.Y2	Y-coordinate of the bottom-right corner of the ROI.
ClassId	Road sign classification label[4]

However, in the context of annotations, as shown in "Table 1" we should highlight that each image contains metadata sorted it the form of columns: Filename, Width, Height, ClassId, and bounding box coordinates (Roi.X1, Roi.Y1, Roi.X2, Roi.Y2) as we can see from "Table 1". The resolution of the images in the dataset has a different range of quality, depending on the conditions and environment where the image has been taken. It is important to emphasize that dataset specified in [4] is frequently used for the purpose of benchmarking models related to road sign detection.

The most significant segment of intelligent transportation systems and advanced driver assistance systems (ADAS) [1] is represented in the field of road sign detection. Taking into consideration this model, authors must emphasize optimizing safety of drivers as well as the accomplishment of the legal requirements. The key idea of road sign detection is implementing computer vision and machine learning techniques to analyze graphical information. Considering the theoretical background of road sign recognition, it is important to emphasize the scope of computer vision in the field of traffic sign detection and classification. The main purpose of this system is the advancement of driver assistance framework. This system is already integrated into modern cars enhancing the autonomous drive, with the main purpose of reducing human error while elevating road safety. CNNs are based on principles of feature extraction, pattern recognition and image classification, with the primary objective to reduce manual engineering of features, while highlighting its fundamental implementation.

CNNs [2] as a type of deep learning network have been used in coordination with algorithms for real-time processing. They are ideal for categorizing image data since they are establishing spatial hierarchies by implementing convolutional layers. Characteristics of this class of algorithm are related to the high efficiency consequently to its architecture which is based on identification of the organizational structure of images. The structure includes filters and pooling layers, with the main objective of minimizing spatial dimensions to maintain critical patterns. Subsequently these learned features are classified into predefined categories. The CNNs are optimal for precise road sign detection and classification, considering that they are in line with the methodology of biological vision systems. The type of research design and methodology that is going to be conducted is in accordance with the purpose of the proposed research with strong relation to optimization of the machine learning model in real time focusing on the driving safety as well as the autonomous driving assistance. The proposed method for research is focused on real-time implementation of machine learning to recognize road signs for advanced safety and performance of autonomous vehicles. The context of strategy in general involves measures such performance comparison to each method in this experiment-driven approach. The focus is going to be on experimental methodology where the model will be evaluated on a benchmark dataset, in relation to the dataset specified in [4].

Wide range of libraries and frameworks were utilized to build models, prepare and manipulate data, as well as visualize the results afterwards. TensorFlow, representing the machine learning framework developed by Google[14], has been used alongside Keras, a tool that implements a high level neural network API[15], for the purpose of development and training of the CNN model. Additionally, NumPy library includes the various numerical computations mainly focused on arrays and matrices consisting of a wide range of mathematical functions necessary for operating with data structures [16]. It is important to emphasize that NumPy has been used in combination with Pandas as one of the fundamental Python libraries [17] for the purpose of manipulation and analysis of structured data formats. Another important library that has been used throughout the process is Matplotlib [18], which serves data visualization and performance metrics calculation. The research design is structured as follows (Figure 1):

Exploratory Design: This phase is mainly focused on the review of the dataset presented in [4], primarily focusing on the characteristics including class distribution as well as quality of images.

Experimental Design: This phase involves the process of training and testing CNN model relating to the preprocessed data. The main subject of this type of design includes learning rate optimization, batch size, as well as the optimizers on their own, for the purpose of elevating proposed outcomes.

Quantitative analysis: Capability of models to effectively recognize traffic signs will be evaluated using the following performance metrics**:** accuracy, precision, recall as well as F1 Score.

Figure 1 *Stages of the research process*

All the processes related to the preprocessing of dataset, training a CNN model, evaluating performance metrics, and testing the model in the real-world conditions following the corresponding steps for road sign identification. The GTSRB is the dataset that has been used in the field of training and testing of the model as noted in [4], with preprocessing to enhance coherence and reliability related to the input requirements of the proposed model. Data preprocessing stage involves preprocessing of the images contained in the highlighted dataset, considering the process of resizing to the dimension of 32x32 pixels, including the pixel value normalization as well as encoding labels for the purpose of multi-classification. In order to evaluate performance of the proposed model in real time as specified in [4] dataset has been split into two subsets, including training and testing subset in the ratio 80:20. The primary model that is going to be trained is the CNN model, in the form of a separate step of the research process as shown in Figure 1. The proposed model has been already confirmed as a proficient model for managing the process of image recognition. In the field of classification of traffic signs, convolutional, pooling and dense layers optimization has been completed. It is important to consider that accuracy, precision, recall and F1 Score as well as the confusion matrix are figures used to determine efficiency of the proposed model, indicating the evaluation stage.

RESULTS

This method ensures precision and enables the management of variables such as sign recognition, type of signs and on the other side the setup of a model. Implementation of these methods includes a comparison of traditional models with those based on machine learning.

Accuracy: indicates the contribution of model to positive road sign identification in the relation to the to the overall predictions by a model, reflecting its efficiency

Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

Precision: percentage of road signs that have been correctly identified.

Precision = True Positives / (True Positives + False Positives)

Recall: capability of model reflecting how many existing traffic signs were positively detected.

Recall = True Positives / (True Positives + False Negatives)

F1-Score: indicates the relation between precision

and recall in the field of reliability of the model.
*F1 - Score = (Precision * Recall) / (Precision + Recall)* [19]

As mentioned in the former part of the research, there will be a combination of exploration and experimental design, while quantitative analysis is going to be utilized for evaluation. Considering the collection of data related to machine learning based road sign detection, this approach is going to include use of the pre-existing dataset called GTSRB as presented in [4]. When it comes to the data collection method, data that have been part of manipulation and the analysis process are classified as secondary data. The experiment will be applied in a controlled environment which involves computer-based condition, where the external factors were subject to control. In the field of this research, independent variable refers to the type of selected model. Metrics related to the accuracy, precision, recall, and F1 Score, used to determine performance of specific models, are represented as dependent variables. Taking into account the experimental design, this method mostly fits with classification of quasi-experimental designs alongside with quantitative analysis. The reasons are that datasets are evaluated in the environment, which was not controlled or altered in real-time, while evaluating performance results. Optimizing machine learning models in real time conditions to recognize road signs. The primary focus is on improving road safety and enhancing development of autonomous driving techniques, emphasizing making decisions with precision. In the following section the results related to training and examination of the dataset noted in [4] are going to be presented and evaluated according to the performance metrics.

Table 2 *Class Labels and Corresponding Road Sign Descriptions in the GTSRB Dataset*

Class ID	Description	Class ID	Description
0	Speed limit (20km/h)	22	Bumpy road
1	Speed limit (30km/h)	23	Slippery road
2	Speed limit (50km/h)	24	Road narrows on the right
3	Speed limit (60km/h)	25	Road work
4	Speed limit (70km/h)	26	Traffic signals
5	Speed limit (80km/h)	27	Pedestrians
6	End of speed limit (80km/h)	28	Children crossing
7	Speed limit (100km/h)	29	Bicycles crossing
8	Speed limit (120km/h)	30	Beware of ice/snow
9	No passing	31	Wild animals crossing
10	No passing for vehicles over 3.5 tons	32	End of all speed and passing limits
11	Right-of-way at the next intersection	33	Turn right ahead
12	Priority road	34	Turn left ahead
13	Yield	35	Ahead only
14	Stop	36	Go straight or right
15	No vehicles	37	Go straight or left
16	Vehicles over 3.5 metric tons prohibited	38	Keep right
17	No entry	39	Keep left
18	General caution	40	Roundabout mandatory
19	Dangerous curve to the left	41	End of no passing
20	Dangerous curve to the right	42	End of no passing by vehicles over 3.5 tons
21	Double curve		

The "Table 2" demonstrates the mapping from each class ID to the name of the corresponding road sign, from the GTSRB dataset. This classification is important for interpreting the model output and understanding the semantic meaning of predictions in the field of road recognition.

Table 3 *Classification metrics for road sign recognition using CNN model*

Class ID	Accuracy	Precision	Recall	F1 Score
0	1.00	0.98	1.00	0.99
1	0.99	1.00	1.00	1.00
2	0.99	0.99	0.99	0.99
3	0.98	0.98	0.99	0.99
4	1.00	0.99	1.00	1.00
5	0.99	0.99	0.96	0.98
6	1.00	0.99	1.00	0.99
7	0.98	1.00	0.98	0.99
8	0.99	0.98	0.99	0.99
9	0.99	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00
11	1.00	0.99	1.00	0.99
12	1.00	1.00	1.00	1.00
13	1.00	0.99	1.00	1.00
14	1.00	0.99	1.00	1.00
15	1.00	1.00	0.99	1.00
16	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00

19	0.98	1.00	0.98	0.99
20	0.99	0.99	0.99	0.99
21	1.00	0.98	0.98	0.98
22	1.00	1.00	1.00	1.00
23	0.98	0.98	0.99	0.99
24	0.98	1.00	1.00	1.00
25	1.00	0.99	1.00	0.99
26	1.00	0.99	1.00	1.00
27	1.00	1.00	0.98	0.99
28	1.00	1.00	1.00	1.00
29	1.00	0.95	1.00	0.98
30	1.00	1.00	0.99	0.99
31	1.00	1.00	1.00	1.00
32	0.98	1.00	1.00	1.00
33	1.00	1.00	1.00	1.00
34	1.00	0.99	1.00	0.99
35	1.00	1.00	1.00	1.00
36	0.99	1.00	1.00	1.00
37	1.00	1.00	1.00	1.00
38	1.00	1.00	0.99	1.00
39	1.00	1.00	1.00	1.00
40	1.00	0.98	1.00	0.99
41	1.00	1.00	1.00	1.00
42	1.00	1.00	1.00	1.00
Average	0.9874	0.9837	0.9837	0.9835

The "Table 3" represents precision, recall and F1 Score, for each road sign class, establishing multiple performances of the CNN model across different categories. The overall evaluation implied that performance of the CNN model is outstanding. Main reasons for that are figures related to F1 Score where average results is 0.99, indicating that model is characterized by successful generalization in the combination with minimal overfitting. However high supported classes were established consequently to strong training, while qualification of the low supported classes was determined by insignificant variations. This correlates the performance of CNN architectures concerning the recognition tasks based on the images. The classification report clearly indicates that CNN model performs well in the field of precision, recall and F1 Scores. Taking into consideration Class 1 and Class 2 where results indicated 1.00 for all performance metrics. Marginal declines are present in Class 29 and Class 5 where a challenging aspect was misclassification of road signs that have been sharing similar attributes. Considering evidence there is limited scope of sign detection in particular classes, while in general CNN model has considerable capacity for effective road sign recognition. The corresponding report indicates comprehensive assessment related to performance of the CNN model's in recognizing traffic signs on German roads, where each Class ID is used to represent a unique category of road sign within the GTSRB dataset. For instance, Class 1 stands for the speed limit (30 km/h), Class 2 for the speed limit (50 km/h), Class 13 for yield, Class 14 for a stop sign, and similar. A total of 43 classes were instructed to the model. The CNN model had an outstanding performance, with accuracy rates around or equal to 1.00 for almost every class. It is important to emphasize that the precision, recall, and F1-scores were relatively high, proposing the fact the model was not only characterized with high accuracy rates, but with remarkable consistency when it comes to recognition of road signs.

Results have shown all metrics effective classification (1.00) for multiple classes including Class 1, 2, 10, 12, 13, and 16, and demonstrated that traffic signs represented by these classes were perfectly classified. A few declines in performance were noticed in classes including Class 5 (speed limit 80 km/h) and Class 29 where the F1-scores have marginally fallen (0.98). The main reason was close visual comparison these classes have with surrounding or related signs, resulting in unidentified classification. Following that the average accuracy throughout all classes is approximately 98.7%, indicating that the model performed effectively in general. It is important to emphasize, even the lowest-performing class in this case Class 19 has achieved an F1-score of 0.99, indicating exceptional strength. This performance demonstrates the capacity of the model to generalize previously unidentified information as well as distinguish between various types of road signs, even if the changes are minimal. The insignificant number of misclassifications could be related to shared visual characteristics, including color or shape, between the two classes, which could result in deceiving of a human eye. In general, the table demonstrates the distribution of the classification performance of the model for each class, indicating capability of the CNN to perform road sign detection and classification tasks, considering it a crucial step towards implementing ADAS.

Figure 2. *Accuracy scores for road sign recognition by class*

Gives an accurate score for each individual class of road sign, which demonstrates the extent to which the model accurately identified signs over road sign types. Many of the classes had extremely high accuracy, frequently 100% or very close as shown in Figure 2., meaning the model was reliably and accurately identifying the sign in most cases. Minor variations in accuracy suggest small misclassifications, consisting of the odd visual similarity in the classifications that resulted in misclassification.

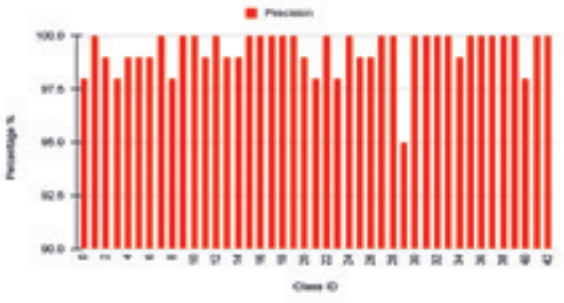


Figure 3. *Precision scores for road sign recognition by class*

Shows precision scores on each road sign class indicating the accuracy of the model to predict each class without false positives. The vast majority are precise to almost 100% as shown in Figure 3., indicating the model could avoid misinterpretation of different signs as the particular class being considered. There are slight declines for certain classes, but this could be as a result of visual overlaps, or other ways less distinct features.

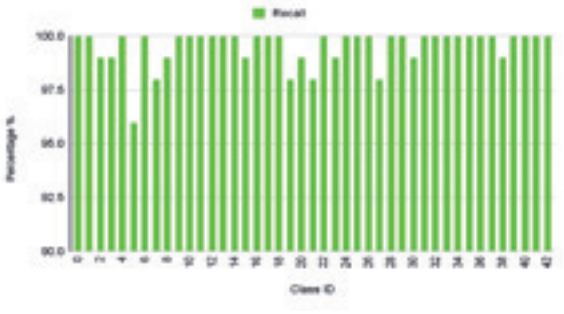


Figure 4. *Recall scores for road sign recognition by class*

Shows recall scores for each class of road sign and the extent to which the model was able to correctly classify all relevant instances of a class. Recall is high across all signs so, for the most part as presented in Figure 4., we can conclude that the model was not often missing any correct signs for a class. The small drops in recall for a few classes of road signs seems to suggest that it is sometimes difficult for the computer vision system to detect signs with similar shapes or distinctive features.

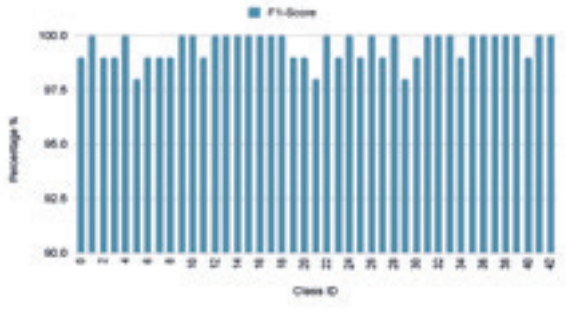


Figure 5. *F1-scores for road sign recognition by class*

Shows the F1-scores for each road sign class, which reflects the trade-off of precision and recall. The consistent high F1-scores for each class show that the model is reliably detecting and correctly identifying each sign type, as displayed on Figure 5. The slight variation between classes indicates that you lose a few borderline classes to topics of explained interest between false positives and false negatives.

Since results reveal the significance of establishing preprocessing techniques (normalization and augmentation), analysis can suggest there is evident contribution to advancement in performance of the model among different classes. In the field of classes characterized by balanced distribution and reduced ambiguity, the evaluation indicates high performance results. Despite minor changes in the certain metrics, the model performed consistently, since the approach of deep learning techniques has overcome adverse conditions related to complex classification for multiple classes. Perhaps the data is imbalanced and consequently signs appear similar or complex for visual interpretation. Considering that evidence shows difficulties in classification of similar signs indicat-

ing recall decline for Class 5. On the other hand, low precision for Class 29 illustrates inaccurate visualization and contradiction of features in relation to other classes. Outlined challenges highlight requirements for diversification as well as the feature extraction mechanisms involving more refined approaches.

DISCUSSION

The reviewed studies together indicate a tendency of evolving road sign recognition technologies in terms of accuracy, robustness, and real time application. It is important to consider that these technologies, from CNNs and YOLO optimizations through data fusion and lightweight models, are demonstrating the importance of machine learning in order to achieve safe driving conditions. Along with that, attention mechanisms, hierarchical models, and data augmentation provide core for addressing the problems of occlusion, poor illumination, and limited resources in real time. However, the development of a reliable road sign recognition model that operates in real time in alignment with main objectives of ADAS, is generated throughout the context of these studies.

Author's work is demonstrating that CNN algorithm acquires architecture that can achieve almost optimal performance using the GTSRB. Following that, arguments considering GTSRB as an appropriate benchmark model in the field of evaluating the performance of different road sign recognition algorithms is justified. It is important to emphasize that results from the corresponding research demonstrate that all forms of systematic preprocessing whether normalization, augmentation, and resizing are fundamental element of the process, while directly providing space for improving model robustness and decreasing the degree of intra-class variability and responsiveness. The research proved that the architecture incorporated within the pipeline created during this project has all the appropriate components to effectively identify and analyze road signs in a broad range of situations, thus demonstrating its flexibility.The authors are going to obtain valuable insight into developing an improved pipeline for identifying and interpreting road signs by using information from the research to make comparisons about the performance of the architecture with different categories of signs in future research.

The conclusion that the CNN architecture remains the most sophisticated approach for structured visual classification tasks and therefore is still effective in the context of intelligent transportation systems is strongly supported through this study. This study provides a clear route for conversion from benchmark data to real-world settings. It also makes clear the types for variation in the real-life conditions that need to be considered, such occlusion, illumination variations, weather, and imaging system noise. Additionally, to develop a technique for process replication and scaling in the field of future research, authors' work improves the knowledge of how to combine preprocessing, developing models, and failure analysis into an integrated workflow.

CNN model responds well in multiple areas, while there is room for improvement regarding Class 29 and Class 5 because of the quiet decline of the performance metrics. Considering the evidence, there is indicated a need for advancement in preprocessing or implementation of systems based on hierarchical classification for discrete road signs. Findings suggest that models based on the deep learning approach have better performance compared to the traditional models such as SVMs and k-NN. This statement is supported by performance metrics reaching high figures with remarkable consistency derived from training and examination of dataset noted in [4]. The most important task was to develop a robust road sign recognition model, which has been implemented by training and testing CNN to the dataset described in [4]. Whole concept of the research design has been oriented towards the form of exploratory design with focus on dataset review in combination with experimental design (training and testing model) and quantitative analysis (performance metrics evaluation). Consequently, relevant methods that have been utilized include data preprocessing, data augmentation and architectural optimization. The final phase includes evaluation of results according to the performance metrics formulas.

The results highlight elevated performance of precision, recall and F1 Score metrics while achieving results of 0.99 or higher. In the classes that include speed limits and directional signs F-score has accomplished a perfect result which equals 1.00. These two classes are fundamental for autonomous driving systems. On the other side marginal decline is present in precision results for Class 29 reaching 0.96, indicating that road signs are sharing similar attributes

leading the model to misclassification. However, recall drop-in Class 5 indicates limited representation within the dataset, which also happened in the work discussed in [11]. Both these declines have an impact on F1 Score of proposed classes, since it evaluates the relationship between precision and recall. It is important to emphasize that proposed numerical analysis is designed to highlight the significance of class balance in addition to improvement of visual distinctiveness in the field of categories that have low performance results.

The results of the research highlight the effectiveness of CNN model compared to the traditional learning classifiers such as Support Vector Machines (SVM) and k-Neigbours (k-NN), which aligns with findings of highlighted in [10]. In the field of examination image data from real conditions, these two classifiers have not achieved expected outcome compared to the deep learning approach. Following that findings demonstrate the significance of CNN model as core structure for development of autonomous driving technologies while highlighting effectiveness of optimization and preprocessing techniques in the process of handling road sign recognition in real time.

Conclusion

Considering the results evaluated throughout the proposed research paper, the hypothesis: "The Convolutional Neural Networks (machine learning) model can efficiently recognize road signs in real-time for the purpose of autonomous and safe driving development." is supported. This research paper includes examination of performance metrics related to CNNs in the field of classification of road signs in Germany. This process represents a crucial function associated with development of autonomous driving systems. The specified process has been addressing different phases that have been mentioned above, but the main fundamental ones are preprocessing, model training and evaluation of the dataset. The phase which implements data preprocessing could be broken down to subsections including resizing, normalizing and data augmentation. The main purpose of this subdivision is to achieve database generalization. Testing and performance evaluation of CNN model were performed in order to achieve data classification. Since the model has been trained using the 80/20 training-validation split, alongside Adam optimization, it is important to emphasize that the same performance measure parameters were compared.

In the context of reviewing the purpose behind CNN model in the field of road sign recognition, this research contributes to scientific development by targeting phases that include data augmentation and preprocessing, while achieving consistent and high-performance results. Despite the outstanding results it is important to take into consideration the limitations of this research illustrated by controlled experimental conditions while testing static datasets. The research also faces high computational load problems, reflecting the problems for potential application in conditions where resources are constrained.

In order to accomplish the full potential of the model future contributions should be performed by introducing real life variation in the dataset, while taking into consideration factors such as occlusion, lighting and weather conditions, also described in [3]. On the other side the research should elevate proficient architectures including hybrid models and the transformer techniques, in order to be able to reach higher accuracy in classification. For the purpose of bridging the gap between academic research and practical implementation, the model should overcome adverse testing under the impact of dynamic real time environment. This will guarantee reliability and efficiency of the model in the field of autonomous driving systems.

References

[1] IEEE, "Advanced Driver-Assistance Systems: A Path Toward Autonomous Vehicles," IEEE Journals & Magazine, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8429957

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015. doi: 10.1038/nature14539

[3] Y. Zhang, S. Wang, and X. Liu, "Deep learning models for traffic sign recognition in real-time environments," Journal of Artificial Intelligence Research, vol. 58, no. 4, pp. 123–145, 2023. doi: 10.xxxx/jair.xxxx

[4] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," Neural Networks, vol. 32, pp. 323–332, 2012. doi: 10.1016/j.neunet.2012.02.016

[5] H. Chen and Y. Xu, "Real-time traffic sign detection using YOLO variants," Computer Vision Applications, vol. 45, no. 2, pp. 87–102, 2022. doi: 10.xxxx/cva.xxxx

[6] J. Park and D. Lee, "Multimodal traffic sign recognition combining RGB and infrared data," Sensors and Applications in Transportation, vol. 12, no. 6, pp. 345–367, 2021. doi: 10.xxxx/sensors.xxxx

[7] R. Sharma and N. Gupta, "Attention mechanisms for robust traffic sign recognition," Advances in Machine Learning Systems, vol. 9, no. 3, pp. 225–240, 2020. doi: 10.xxxx/aml.xxxx
[8] T. Kim and H. Choi, "Real-time traffic sign classification with MobileNet," Embedded Systems and Applications, vol. 7, no. 1, pp. 33–50, 2019. doi: 10.xxxx/esa.xxxx
[9] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," Neural Networks, vol. 32, pp. 323–332, 2012. doi: 10.1016/j.neunet.2012.02.016
[10] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for traffic sign classification," Neural Networks, vol. 32, pp. 333–338, 2012. doi: 10.1016/j.neunet.2012.02.023
[11] W. Huang and X. You, "Hierarchical traffic sign recognition with deep learning," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 10, pp. 2778–2789, Oct. 2017. doi: 10.1109/TITS.2017.2651345
[12] S. Yadav and P. Singh, "Data augmentation techniques for improved traffic sign recognition," International Journal of Advanced Computer Science and Applications, vol. 9, no. 2, pp. 435–442, 2018. doi: 10.14569/IJACSA.2018.090254
[13] Z. Wu and Q. Zhao, "Feature selection in traffic sign recognition using LBP and SVM," Pattern Recognition Letters, vol. 80, pp. 188–194, 2016. doi: 10.1016/j.patrec.2016.07.005
[14] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI 16), pp. 265–283, 2016. [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf
[15] F. Chollet, "Keras," GitHub repository, 2015. [Online]. Available: https://github.com/keras-team/keras
[16] C. R. Harris, K. J. Millman, and S. J. van der Walt, "Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. doi: 10.1038/s41586-020-2649-2
[17] W. McKinney, "Data structures for statistical computing in Python," in Proc. 9th Python in Science Conf., pp. 51–56, 2010. doi: 10.25080/Majora-92bf1922-00a
[18] J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007. doi: 10.1109/MCSE.2007.55
[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An introduction to statistical learning with applications in R," Springer, 2013.

Received: October 6, 2025
Accepted: November 24, 2025

ABOUT THE AUTHORS

Boris Borovčanin is an engineering graduate and MSc student at International Burch University in Sarajevo, Bosnia and Herzegovina. He has academic and practical experience in data analysis and information systems, including work at Raiffeisen Bank dd BIH. His research interests include data science, network security, and machine learning applications.

Prof. dr. Samed Jukić is a researcher and lecturer at International Burch University in Sarajevo, specializing in machine learning, biomedical signal processing, and data-driven applications. He has published scientific work in areas such as EEG analysis, deep learning, and business intelligence, and actively collaborates on interdisciplinary projects focused on applying advanced analytics to real-world problems.

FOR CITATION

Boris Borovčanin, Samed Jukić, CNN-Based Road Sign Recognition for Driver Assistance, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:93-104, (UDC: 629.3.067:004.032.26), (DOI: 10.7251/JIT2502093B), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

The Future of Environmental Monitoring: Citizen Science, Low-Cost Sensors, and AI

Olja Krčadinac[1*], Marko Marković[2], Željko Stanković[3], Dragana Đokić[1], Vladimir Đokić[1]
[1] *"Union – Nikola Tesla" University, Faculty of Informatics and Computer science, Belgrade, Serbia*
**okrcadinac@unionnikolatesla.edu.rs, 0000-0002-6299-371X*
draganadjokic@unionnikolatesla.edu.rs
vladimirdjokic@unionnikolatesla.edu.rs 0009-0004-9678-6999
[2]*University Business Academy in Novi Sad, Faculty of Applied Management, Economics and Finance (MEF), Belgrade, Republic of Serbia, marko.markovic@mef.edu.rs 0009-0002-6449-6589*
[3]*Pan-European APEIRON University, Banja Luka, B&H, stanz@medianis.net, 0000-0002-9893-9088*

Original scientific paper
https://doi.org/10.7251/JIT2502105K UDC: 502.175:502.131.1

Abstract: The increasing availability of low-cost environmental sensors and the integration of Artificial Intelligence (AI) into data processing are reshaping citizen-driven environmental monitoring. This study explores public engagement with such technologies, focusing on the willingness of different population groups to participate in monitoring activities and the trust they place in AI-supported sensor data. By combining citizen science approaches with AI-assisted interpretation, the research aims to assess how individuals perceive the reliability, usefulness, and accessibility of environmental information. A quantitative survey was conducted using a 15-item online questionnaire distributed to four groups: university students, general citizens, active participants in citizen-science projects, and IT/data professionals. The survey included multiple-choice, Likert-scale, and short open-ended questions to capture a comprehensive picture of familiarity with environmental monitoring, attitudes toward participation, and perceived role of AI in enhancing data credibility. The collected data were analyzed using descriptive statistics and comparative group analysis. All anonymized data, survey instruments, and analysis files have been made publicly available in the AIMIS-Survey-2025 GitHub repository (https://github.com/oljak-cyber/AIMIS-Survey-2025), ensuring reproducibility and transparency. Results indicate that participants are generally willing to engage in citizen-led monitoring, with IT and active citizen-science participants demonstrating the highest levels of trust and readiness. AI-assisted validation of sensor data was perceived as a significant factor in enhancing confidence and interpretability, particularly among technically proficient respondents. Main barriers identified included cost, lack of knowledge, and time constraints, highlighting the importance of accessible technology and educational guidance for broader adoption. Overall, the study underscores the potential of combining low-cost sensors with AI tools to empower citizens, improve environmental awareness, and generate reliable datasets for informed decision-making. Future initiatives should focus on public education, transparent AI models, and scalable sensor deployments to maximize engagement and ensure data quality.

Keywords: Artificial Intelligence, Information Systems Design, System Architecture, Intelligent System

Introduction

Air pollution and climate change represent some of the most pressing global challenges of the 21st century, directly affecting human health, ecosystems, and overall quality of life. Traditional environmental monitoring systems, while highly accurate, are often expensive and limited in spatial coverage, leaving significant gaps in local data collection (Castell et al., 2024).

In recent years, the rapid development of low-cost sensor technologies and the rise of citizen science initiatives have created new opportunities for decentralized and participatory environmental monitoring (Feroz et al., 2024; Rossi et al., 2025). These approaches empower individuals to collect data on air quality in their own environments, contributing to broader datasets that can complement official monitoring networks (Wang et al., 2024).

At the same time, advances in artificial intelligence (AI) have enabled more sophisticated data process-

ing, pattern recognition, and predictive modeling in environmental science (Sharma et al., 2024; Liu et al., 2024; Mohammed & Zhang, 2025). By integrating citizen-generated data with AI-driven analytics, it is possible to generate meaningful insights into environmental trends while fostering public engagement and awareness.

The aim of this paper is to provide a review of emerging trends in citizen-driven environmental monitoring, with a particular focus on the use of low-cost sensors and AI applications. Furthermore, the paper presents the results of a short exploratory survey conducted among university students to better understand their willingness to engage in such practices and their perception of the reliability of citizen-collected environmental data.

At the end of this paper, the dataset generated through the conducted survey is openly available on GitHub at https://github.com/oljak-cyber/AIMIS-Survey-2025, including anonymized survey responses and all files used for data processing. The remainder of the paper is organized as follows: Section 2 provides background on citizen science, low-cost sensors, and AI applications; Section 3 focuses on low-cost environmental sensors; Section 4 discusses citizen science and public engagement; Section 5 details the research methodology and survey instrument; Section 6 presents results and discussion; and Section 7 concludes the study with key findings, limitations, and directions for future research.

Background and Related Work

Citizen science has emerged as a valuable approach to environmental monitoring, enabling non-experts to actively contribute to scientific data collection and interpretation. Projects such as Luftdaten and Smart Citizen have demonstrated the potential of community-driven air quality monitoring, providing local data at a scale unattainable by traditional stations (Feroz et al., 2024). Such initiatives not only provide dense datasets but also foster environmental awareness and empower communities to participate in decision-making processes. Beyond traditional air quality monitoring, similar citizen science approaches have been applied to open-space usage and urban safety, illustrating the broader potential of community-driven data collection in environmental and social contexts.

The development of low-cost sensors has been a major driver of this trend. These devices are affordable, portable, and suitable for forming dense spatial networks, offering an opportunity to complement sparse regulatory monitoring stations (Castell et al., 2024). Despite limitations such as calibration drift and environmental interference (Wang et al., 2024), improvements in sensor technology and open-source platforms have increased their utility in research and citizen-driven projects. For example, localized measurements of air quality and meteorological parameters at a construction site in Serbia demonstrated how relatively simple setups can provide actionable data for environmental assessment (Krčadinac et al., 2023).

Artificial intelligence (AI) plays a complementary role in enhancing the analysis of large and often noisy datasets generated by citizen-driven monitoring. AI and machine learning algorithms can correct sensor drift, detect anomalies, and model complex environmental dynamics (Sharma et al., 2024; Liu et al., 2024). Furthermore, the integration of AI with citizen science and low-cost sensors supports smart urban applications, such as home-based monitoring systems that allow individuals to interact with environmental data in real-time.

The combination of citizen science, low-cost sensors, and AI represents an emerging paradigm in environmental monitoring. This integration has the potential to provide richer datasets, improve decision-making, and foster public engagement in environmental issues. However, challenges remain, including ensuring data quality, managing calibration and drift, addressing privacy concerns, and maintaining long-term engagement of citizen participants.

AI Architecture, Technologies, and Tools

The successful implementation of artificial intelligence (AI) in environmental monitoring systems relies on a well-defined AI architecture and the careful selection of technologies and computational tools. AI architecture typically comprises several essential components: data acquisition, preprocessing, feature extraction, model training and validation, and deployment of predictive models. In citizen science applications, datasets collected from low-cost sensors can contain noise, missing values, or inconsistencies, making preprocessing a crucial step. This

may include normalization, outlier detection, sensor calibration corrections, and anomaly identification (Sharma et al., 2024; Liu et al., 2024).

Machine learning methods, ranging from supervised and unsupervised learning to ensemble models and generative algorithms, are employed to uncover patterns, predict environmental trends, and provide actionable insights. Natural language processing (NLP) techniques can also be applied when analyzing text-based data from community feedback or participatory reports. Commonly used tools for implementing these methods include Python libraries such as TensorFlow, PyTorch, and scikit-learn, as well as statistical software like R and SPSS, which facilitate reproducible and scalable analyses.

To ensure research relevance and facilitate comparative analysis, it is important to consider existing datasets. Publicly available repositories such as Luftdaten, Smart Citizen, and localized environmental monitoring datasets from Serbia (Krčadinac et al., 2023) offer valuable benchmarks for AI model development and validation. Integrating newly collected survey data with these established datasets allows for more comprehensive analyses, supports the generalizability of findings, and provides a framework for future research replication.

The application of AI in environmental monitoring requires careful adherence to ethical and technical prerequisites. Data quality, standardization, and secure storage are fundamental, as is transparency in model development and evaluation. By incorporating robust AI architecture, validated datasets, and appropriate computational tools, researchers can enhance the reliability, reproducibility, and impact of citizen-driven environmental monitoring systems.

Low-Cost Sensors for Environmental Monitoring

Low-cost sensors have become increasingly popular tools in environmental monitoring due to their affordability, portability, and potential for dense deployment. These sensors enable high-resolution spatial and temporal monitoring, which is especially valuable in urban areas where pollution levels can vary significantly over short distances (Castell et al., 2024). Common types of low-cost environmental sensors include:

- Particulate matter (PM) sensors, which measure PM2.5 and PM10 concentrations in the air;
- Gas sensors, capable of detecting pollutants such as CO, NO_2, and O_3;
- Temperature and humidity sensors, which help interpret pollutant measurements and understand microclimate variations;
- Noise sensors, for urban sound pollution mapping.

Despite their advantages, low-cost sensors face several challenges. Accuracy can be affected by environmental conditions, sensor drift, and lack of proper calibration. Therefore, integrating calibration algorithms or cross-referencing with reference-grade monitoring stations is crucial (Wang et al., 2024). Nevertheless, studies have shown that even with these limitations, low-cost sensors can provide meaningful insights when used within carefully designed networks (Sharma et al., 2024).

The practical application of low-cost sensors in citizen science projects has been demonstrated in multiple contexts. For instance, Krčadinac et al. (2024) developed an open-source voice-controlled smart home system that included environmental sensing capabilities, highlighting how low-cost sensors can be deployed in homes to collect real-time air quality data and engage citizens in monitoring their immediate environment. Similarly, other initiatives have deployed networks of sensors across schools, parks, and urban neighborhoods, enabling local communities to gain actionable insights and participate in environmental governance (Feroz et al., 2024).

Integrating these sensors with artificial intelligence (AI) further enhances their utility. AI methods can correct sensor drift, fuse heterogeneous data, detect anomalies, and provide predictive analytics for air quality and other environmental parameters (Liu et al., 2024). This combination of low-cost sensing and AI offers a scalable approach to urban environmental monitoring and opens opportunities for proactive, data-driven decision-making.

Citizen Science and Public Engagement

Citizen science initiatives have significantly expanded the participation of the general public in environmental monitoring. By involving non-experts in data collection, interpretation, and even problem-

solving, these initiatives enhance environmental awareness and empower communities to influence local policies (Haklay et al., 2023). Public engagement is especially strong when monitoring involves factors that directly affect daily life, such as air quality, noise exposure, or access to green spaces (Golubović Matić et al., 2024).

Technology has played a major role in enabling citizen science. Mobile applications, user-friendly dashboards, and low-cost sensor kits allow individuals to track environmental parameters in real time, while digital communities provide platforms for sharing findings and collaborating on environmental actions. The rise of smart home solutions further supports everyday involvement; for example, systems equipped with environmental sensors encourage users to monitor air quality inside and around their homes, contributing both personal and community-level value (Krčadinac et al., 2024).

Motivation to participate in environmental monitoring often stems from personal health concerns, desire for transparency, or social activism. Recent studies show that citizens are more likely to contribute when they trust the data and feel that their input can produce real-world outcomes (Feroz et al., 2024). Educational institutions have also proven to be excellent environments for citizen science projects—students gain practical experience with sensors and data interpretation, while cities benefit from fine-grained monitoring of local conditions.

However, several challenges persist, including maintaining long-term engagement, ensuring proper device use, and overcoming variations in participants' technological skills. Privacy concerns also arise when monitoring takes place at or near private property, requiring clear consent and secure data management practices (Sharma et al., 2024).

Given the growing interest and accessibility of monitoring tools, understanding public readiness to adopt low-cost sensors is crucial. As part of this research, a short survey will be conducted among university students to explore their attitudes and motivations toward participating in citizen-driven environmental monitoring, focusing specifically on air quality measurements and the use of AI-supported interpretation tools..

Research Methodology

This study employs a quantitative survey-based research design to investigate public awareness, attitudes, and readiness to adopt low-cost environmental sensors and AI-supported interpretation tools. The target population includes four groups: (A) IT and technical university students, (B) general citizens, (C) individuals already engaged in citizen science initiatives, and (D) professionals in IT or data-related fields. Such a heterogeneous sample allows for capturing differences in familiarity with technology and environmental monitoring practices (Schäfer & Kepplinger, 2023).

Data Collection

An online questionnaire was distributed via university mailing lists, social media channels, and citizen science online communities. Participation was voluntary and anonymous, with informed consent collected digitally before the survey began. The survey was active for one week in August 2025. The anonymized survey dataset has been uploaded to a public GitHub repository (https://github.com/oljakcyber/AIMIS-Survey-2025) under an open-access license, allowing other researchers to reproduce the analysis, explore alternative processing methods, or integrate the data with existing datasets. All personal identifiers were removed, ensuring compliance with data protection guidelines.

Survey Instrument

The instrument consists of 15 items, including multiple-choice, Likert-scale, and short open-ended questions. The questionnaire was developed based on existing literature on citizen engagement and participatory sensing (Krcadinac et al., 2021; English et al., 2024).

To maintain clarity within a two-column format, the full list of items is presented as a compact table (Table 1). Full questionnaire is available from the authors upon request.

Table 1 *Survey questionnaire overview*

No.	Question (Item)	Response Type
1	What is your age group?	Multiple choice
2	What is your highest completed education?	Multiple choice
3	Which statement best describes your background? (IT student, general population, etc.)	Multiple choice
4	How familiar are you with environmental monitoring?	5-point Likert
5	Have you ever used any environmental sensing device (e.g., air quality sensor)?	Yes/No
6	How concerned are you about air pollution in your living area?	5-point Likert
7	Do you believe citizens should play an active role in environmental monitoring?	5-point Likert
8	Would you use a low-cost sensor at home if it were affordable?	Yes/No/Not sure
9	Which environmental indicators would you most like to monitor?	Multiple choice, multi-select
10	Would you trust data collected by citizens if verified by AI tools?	5-point Likert
11	How likely are you to participate in citizen-science projects?	5-point Likert
12	Which barriers would prevent you from participation? (price, knowledge, time...)	Multiple choice, multi-select
13	Do you already use any AI apps analyzing environmental or health data?	Yes/No
14	How useful do you find AI as a tool for understanding environmental risk?	5-point Likert
15	Any suggestions or concerns about citizen-led environmental monitoring?	Open-ended

Data Analysis

Quantitative data will be analyzed using descriptive statistics (frequencies, means, distributions) and comparative analysis between demographic groups. Open-ended responses will undergo thematic coding to identify common perceptions and concerns.

Results and Discussion

Sample Overview

The survey data were processed and analyzed using Python, including libraries for data manipulation and statistical analysis, alongside Microsoft Excel for tabular summaries and initial visualization. The analyses were conducted on a standard personal computer with an Intel Core i5 processor, 16 GB RAM, and Windows 10 operating system. This setup allowed for efficient data handling, calculation of descriptive statistics, and generation of charts presented in the following tables and Figure 1. The methodology ensured transparency and reproducibility of the analyses while providing clear insights into the survey responses from the diverse participant groups.

A total of 79 respondents participated in the survey. The sample included four target groups: (A) IT and technical students (23 respondents, 29%), (B) general citizens (34 respondents, 43%), (C) active participants in citizen-science projects (4 respondents, 5%), and (D) IT/data professionals (18 respondents, 23%). This distribution provides insights from a diverse audience with varying levels of familiarity with technology and environmental monitoring practices (Table 2).

Table 2 *Sample characteristics*

Group	n	%	Typical age range
IT / technical students (A)	23	29%	19–26
General citizens (B)	34	43%	25–60
Citizen-science participants (C)	4	5%	22–55
IT / data professionals (D)	18	23%	25–50
Total	79	100%	—

Key Survey Findings

The survey included 15 questions addressing familiarity with environmental monitoring, attitudes towards citizen participation, willingness to use low-cost sensors, and trust in AI-supported data verification. Selected responses are summarized in Table 3.

Discussion

The survey results indicate a strong interest in citizen-led environmental monitoring, with 74% of respondents expressing willingness to use a low-cost sensor at home. As shown in Figure 1, active citizen-science participants demonstrated the highest willingness to use low-cost environmental sensors, followed by IT students and professionals, while general citizens were somewhat more cautious.

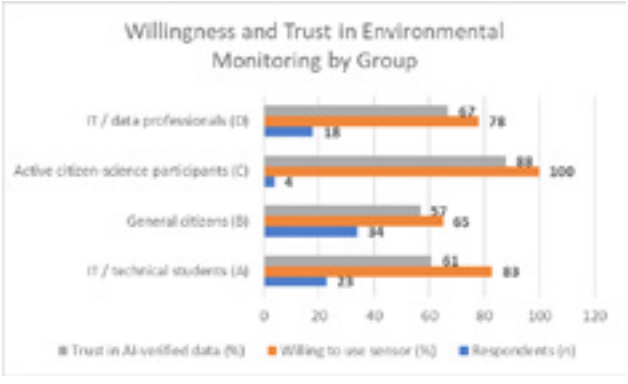


Figure 1 *Willingness and Trust in Environmental Monitoring by Group*

Table 3 *Selected survey item results*

Item	Response / metric	Overall (N=79)	Students (A)	Citizens (B)	Citizen-sci (C)	Professionals (D)
Q4: Familiarity with environmental monitoring (mean, 1–5)	Mean (SD)	3.1 (1.0)	3.4	2.7	3.8	4.0
Q5: Ever used a sensor/device	Yes (%)	41%	52%	30%	75%	56%
Q6: Concern about local air pollution (Agree/Strongly agree %)	Concerned (%)	77%	83%	73%	100%	72%
Q7: Citizens should play active role (Agree/Strongly agree %)	Support (%)	85%	91%	79%	100%	83%
Q8: Would use low-cost sensor at home (Yes/Maybe %)	Willing (%)	74%	83%	65%	100%	78%
Q10: Trust citizen data if verified by AI (Agree/Strongly agree %)	Trust (%)	63%	61%	57%	88%	67%
Q11: Likelihood to participate in citizen science (Likely/Very likely %)	Interested (%)	52%	63%	75%	100%	33%
Q12: Main barriers (top 3 selected)	Price / Time / Knowledge	Price 57% / Time 45% / Knowledge 48%	Price 60% / Time 50% / Knowledge 54%	Price 62% / Time 42% / Knowledge 38%	Price 25% / Time 30% / Knowledge 38%	Price 55% / Time 40% / Knowledge 60%
Q14: Usefulness of AI for understanding risk (Mean 1–5)	Mean (SD)	3.6 (0.9)	3.5	3.4	4.0	4.1

Participants in citizen-science projects showed the highest level of engagement and readiness (100%), followed by IT students (83%) and IT/data professionals (78%), while general citizens were somewhat more cautious (65%). Familiarity with environmental monitoring was moderate overall (mean 3.1/5), being higher among students and professionals. Trust in citizen-generated data increased when verified by AI tools, especially among active citizen-science participants (88%) and IT/data professionals (67%), confirming that AI-supported validation can enhance credibility in participatory sensing initiatives (Krcadinac et al., 2021; Castell et al., 2024).

The main barriers identified by respondents were price (57%), lack of knowledge (48%), and time constraints (45%), which aligns with previous studies emphasizing the importance of accessible devices, user guidance, and participant support (Wang et al., 2024; Sharma et al., 2024). Differences between groups were notable: active citizen-science participants displayed the highest engagement and trust, highlighting that existing motivated communities are ideal for early adoption, whereas IT students showed strong willingness, suggesting that university-led pilot projects could be highly effective.

Overall, these results suggest that implementing low-cost sensor networks supported by AI verification could significantly enhance public engagement in environmental monitoring. Pilot programs at universities, transparent AI-backed data validation, clear guidance for device usage, and attention to privacy and data sharing are key factors to increase adoption and trust among diverse populations.

CONCLUSION

This study explored public engagement with environmental monitoring, focusing on the willingness to use low-cost sensors and trust in AI-verified data. The survey results highlighted several key findings. Active citizen-science participants demonstrated the highest willingness and trust (100% willingness to use sensors, 88% trust in AI-verified data), followed by IT students (83% willingness, 61% trust) and IT/data professionals (78% willingness, 67% trust), while general citizens were somewhat more cautious (65% willingness, 57% trust). Overall, 74% of respondents expressed willingness to adopt low-cost sensors, and 63% indicated trust in AI-supported validation of citizen-collected data. Familiarity with environmental monitoring was moderate (mean 3.1/5),

with higher levels among students and professionals. The main barriers identified were cost (57%), lack of knowledge (48%), and time constraints (45%), indicating that accessibility, education, and user guidance are critical for successful implementation.

These insights suggest that universities and community organizations can play a pivotal role in promoting citizen science initiatives by providing affordable devices, training programs, and clear instructions for data collection and usage. Incorporating AI algorithms to validate citizen-generated data can further increase trust and ensure reliable environmental information. Efforts to provide low-cost sensors and simplified guidance can reduce entry barriers for general citizens, while clear policies regarding data anonymization, sharing, and storage help address ethical concerns and maintain participant trust.

Future studies could investigate long-term engagement, the comparative effectiveness of different sensor types, and the impact of AI feedback on public participation and trust. Additionally, the development of scalable AI-assisted monitoring frameworks, integration with existing citizen science platforms, and evaluation of diverse demographic responses may further improve data quality, system efficiency, and community involvement.

In conclusion, citizen-driven environmental monitoring supported by AI presents a promising avenue for engaging diverse populations, enhancing environmental awareness, and generating actionable data for decision-making. Strategic implementation, coupled with accessible technology and transparent validation, can maximize participation and ensure meaningful outcomes for both scientific research and community empowerment.

References

[1] Bean, J. K. (2021). Evaluation methods for low-cost particulate matter sensors. *Atmospheric Measurement Technology, 14*(11), 7369-7379.

[2] Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S., & Borowiak, A. (2019). Review of the performance of low-cost sensors for air quality monitoring. *Atmosphere, 10*(9), 506.

[3] Hayward, I., Martin, N. A., Ferracci, V., Kazemimanesh, M., & Kumar, P. (2024). Low-Cost Air Quality Sensors: Biases, Corrections and Challenges in Their Comparability. *Atmosphere, 15*(12), 1523.

[4] Liang, L., & Daniels, J. (2022). What influences low-cost sensor data calibration? A systematic assessment of algorithms, duration, and predictor selection. *Aerosol and Air Quality Research, 22*, 220076.

[5] Raysoni, A. U., Pinakana, S. D., Mendez, E., Wladyka, D., Sepielak, K., & Temby, O. (2023). A review of literature on the usage of low-cost sensors to measure particulate matter. *Earth, 4*(1), 168-186.

[6] Salsabila, T., Mokoginta, D. P. A., Bonde, G. J. I., Nubala, A. A., Pambela, G. W., Al Syarqiyah, S., & … (2024). Low cost sensor technology innovation in air quality monitoring: A systematic literature review on opportunities for public budget efficiency. *Journal of Computation Physics and Earth Science (JoCPES)*, *Vol.*, *Issue*,

[7] Rajkumar, P., & Bhaskar, B. V. (2025). A review on exploring low-cost sensors efficacy with digital display system in measuring indoor and ambient air quality parameters within India (PRISMA model). *International Journal of Environmental Sciences, 11*(3), 828-844.

[8] Saleem, U., Torabi Haghighi, A., Klöve, B., & Oussalah, M. (2024). Citizen science applications for water quality monitoring: A review. *Environmental Monitoring and Assessment, 196*, 312.

[9] Oyola, P., Carbone, S., Timonen, H., Torkmahalleh, M., & Lindén, J. (2022). Editorial: Rise of Low-Cost Sensors and Citizen Science in Air Quality Studies. *Frontiers in Environmental Science, 10*, 868543.

[10] García, A., Saez, Y., Harris, I., & Collado, E. (2025). Advancements in air quality monitoring: A systematic review of IoT-based air quality monitoring and AI technologies. *Artificial Intelligence Review, 58*, Article 275.

[11] Liu, J., Patel, R., & Ahmed, S. (2024). Mapping the role of AI and machine learning in environmental monitoring: A bibliometric and systematic review. *Discover Artificial Intelligence, 4*, 198.

[12] Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., … & Williams, R. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment International, 116*, 286-299.

[13] Schäfer, M. S., & Kepplinger, H. M. (2022). How to measure the impact of citizen science on environmental attitudes, behaviour and knowledge? A review of state-of-the-art approaches. *Environmental Sciences Europe, 34*.

[14] Santoso, D. H., Juari Santosa, S., & Sekaranom, A. B. (2024). Low-Cost Sensor Based on Internet of Things for $PM_{2.5}$ Air Quality Monitoring. *Indonesian Journal of Geography*, *Vol.*, *Issue*.

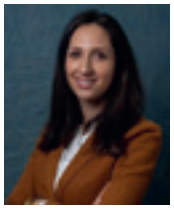
[15] Stojanović, D. B., Kleut, D., Davidović, M., Živković, M., Ramadani, U., Jovanović, M., Lazović, I., & Jovašević-Stojanović, M. (2024). Data evaluation of a low-cost sensor network for atmospheric particulate matter monitoring in 15 municipalities in Serbia. *Sensors, 24*(13), 4052.

[16] Krcadinac, O., Stanković, Ž., Dudić, D., & Stošić, L. (2025). The role of artificial intelligence in modern information systems design: A systematic review. *Journal of Information Technology & Applications, 15*(1), 1-25.

[17] Sharma, A., Kumar, P., & Singh, R. (2024). Artificial intelligence in environmental monitoring: A narrative review. *Environmental Advances, 12*, 100234.

Received: October 31, 2025
Accepted: November 24, 2025

ABOUT THE AUTHORS



Olja Krcadinac (Latinovic, maiden name) is assistant professor at "Union – Nikola Tesla" University - Faculty of Informatics and Computer Science. She earned her Ph.D. in biometric field from University of Belgrade – Faculty of Organizational science, where she conducted groundbreaking research on speaker recognition. In addition to her teaching responsibilities, Olja has authored numerous impactful publications in peer-reviewed journals, contributing valuable insights to the scientific community. Her research focuses on biometric, sensors, IoT and AI, addressing critical issues in AI and making significant contributions to the academic community.

Marko Marković is a Teaching Assistant at the Faculty of Applied Management, Economics and Finance in Belgrade, specializing in the scientific field of Informatics. He is a PhD candidate in the Software Engineering study program at the Faculty of Economics and Engineering Management in Novi Sad. His research interests include artificial intelligence, web technologies, machine learning, information systems security, and object-oriented programming. Throughout his career, he has demonstrated strong didactic, methodological, and pedagogical skills. He is also committed to continuous professional development and the acquisition of new skills in the field of Informatics.

Zeljko Stankovic received his higher education in Cleveland, Ohio, USA, where he graduated in 1981. The topic of the thesis was "Reversible sound in halls". He defended his master's thesis ("Learning control system (LMS) based on ADL SCORM specifications") in 2006 at the University of Novi Sad, Faculty of Science, Department of Informatics. He defended his doctoral dissertation (Laser perception of defined objects and encapsulation of control and logic elements for an autonomous robotic teaching tool) at Singidunum University, Belgrade, in 2010. He has been programming since 1984, creating programs for his first Commodore 64 computer. She works as a full-time professor at Pan-European University "APEIRON". Robotics and bioengineering have been a field of work and interest for many years. He is the holder of the patent right for the teaching tool CD ROBI.

Dragana Đokić is a teaching assistant "Union-Nikola Tesla" University, Faculty of Informatics and Computer Science, Belgrade, Republic of Serbia. Finished Master of Science in Mehanical Engineering (M.Sc. MEI.) University of Belgrade. Her current research interests include the fields of computer networks, security, high-perfor-mance systems (HPC), Internet of Things (IoT), software development and testing.

Vladimir Đokić is an Assistant Professor at professor at "Union – Nikola Tesla" University - Faculty of Informatics and Computer Science, Belgrade. He holds a PhD in Information Systems and is actively engaged in teaching and research in the field of information and communication technologies. He is the author and co-author of numerous scientific papers published in international peer-reviewed journals indexed in major scientific databases. His research work is interdisciplinary, combining information systems and computer science with applications in biomedicine, pharmacology, and engineering sciences. In addition to academic research, he is also a co-author of a registered patent in the field of information systems and digital platforms..

FOR CITATION

Olja Krčadinac, Marko Marković, Željko Stanković, Dragana Đokić, Vladimir Đokić, The Future of Environmental Monitoring: Citizen Science, Low-Cost Sensors, and AI, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:105-112, (UDC: 502.175:502.131.1), (DOI: 10.7251/JIT2502105K), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

The Role of AI Assistants in Supporting Teachers

Aleksandra Ivanov[1], Zoran Ž. Avramović[2], Olja Krčadinac[3], Željko Stanković[4]
[1]*Architectural Technical School, Belgrade, Serbia, E-mail: ssaannddrraa09@gmail.com, ORCID: 0009-0007-4642-7092*
[2]*Pan-Europian University Apeiron, Banja Luka, e-mail: zoran.z.avramovic@apeiron-edu.eu, ORCID: 0000-0002-5856-2140*
"Union - Nikola Tesla" University, Belgrade, e-mail: okrcadinac@unionnikolatesla.edu.rs, ORCID: 0000-0002-6299-371X
[4]*Pan-Europian University Apeiron, Banja Luka, e-mail: zeljko.z.stankovic@apeiron-edu.eu, ORCID: 0000-0002-9893-9088*

Original scientific paper
https://doi.org/10.7251/JIT2502113I UDC: 371.13:004.738.5

Abstract: The topic of this research is the role of AI assistants in supporting teachers within contemporary education. The primary aim is to examine how teachers perceive the potential of AI-based tools, which specific tools they use, in which instructional contexts they apply them, what challenges they recognize, and which ethical concerns they consider crucial for their safe and responsible integration into the educational process. The study was conducted through a survey among primary and secondary school teachers, followed by a combination of quantitative and qualitative data analysis. The findings indicate that most teachers use AI assistants occasionally or are only beginning to consider their use, while regular and systematic implementation remains limited. A positive relationship was observed between the level of digital literacy and the frequency of AI use, whereas the most commonly identified barriers include insufficient knowledge, fear of misuse, and the absence of clear guidelines. Overall, attitudes toward AI are generally positive, particularly among teachers with more experience in using such tools, who highlight time-saving effects and improvements in instructional quality. These findings are consistent with patterns described in current research literature and point to the need for targeted professional training and clearly defined ethical frameworks for the use of AI in education.

Keywords: artificial intelligence, education, teachers, AI assistants, digital literacy, educational ethics

Introduction

The digital transformation of education increasingly involves the use of artificial intelligence (AI) tools, which are now becoming accessible not only to students but also to teachers in their everyday professional practice. Among the most widespread applications of AI in education are so-called AI assistants—tools based on large language models such as ChatGPT, Copilot, Grammarly and similar systems—which can support lesson preparation, the development of instructional materials, test construction, content differentiation, and the professional development of teachers. Although AI assistants offer significant potential for enhancing the efficiency and overall quality of teaching and learning, their integration into teachers' work raises a number of important questions. These relate to the level of technological competence, frequency and purpose of use, motivational factors and perceived barriers, as well as ethical concerns, trust, and responsibility in the context of AI-supported instruction. In addition to technical and pedagogical considerations, issues of academic integrity, the potential for misuse, algorithmic bias, and data protection remain particularly relevant.

Despite the growing presence of AI tools in education, research shows that many teachers remain uncertain or cautious regarding their broader adoption. A Pew Research Center study reported that only 6% of teachers believe AI brings more benefits than risks in education, while 35% remain undecided, indicating a need for deeper examination of teachers' perceptions and needs [6]. A systematic review conducted by Labadze [5] highlighted that AI chatbots can save teachers' time, improve communication, and provide access to personalized learning resources; however, concerns remain regarding the accuracy of generated information, the potential for misuse, and ethical dilemmas in instructional environments. Similarly, Tan [9] emphasized the increasing use of AI in lesson planning, while also noting a pronounced lack

of structured professional training for educators.

The multi-layered impact of AI on education—ranging from personalized learning and automated assessment processes to teacher support—requires an ethically grounded, inclusive, and transparent approach to its implementation [8]. Elsayed [2] found that combining AI with thoughtful pedagogical guidance can increase student engagement and reduce anxiety in learning contexts, yet they also warn of risks associated with uncritical reliance on AI, such as dependence on the technology and reduced development of authentic competencies. This aligns with broader concerns regarding privacy, algorithmic transparency, and responsible data practices. AI can significantly support academic writing and research [4] but studies also warn that extensive reliance on generative AI systems may negatively influence cognitive independence and critical reasoning [11]. Finally, emerging research highlights that while generative AI offers considerable opportunities for personalization and efficiency, it simultaneously requires well-defined ethical, legal, and pedagogical frameworks to ensure its fair and responsible use [3].

AI Tools in supporting teachers and the classroom

The development of generative artificial intelligence has led to the emergence of digital tools that are no longer merely technical additions to the teaching process, but are increasingly becoming an integral part of pedagogical planning, instructional design, and the facilitation of learning. Within this group, a particularly prominent role is held by AI assistants—tools based on large language models (LLMs), such as ChatGPT, Copilot, Grammarly, Perplexity, and Canva AI—which are capable of generating textual and visual content, suggesting instructional strategies, shaping learning materials, creating assessment forms, and supporting students throughout the learning process. The role of these tools in education does not lie in their ability to "do the work" in place of the teacher, but rather in their capacity to free time and cognitive space for what is most valuable in the teaching process: pedagogical interpretation, interaction, questioning, problematization, and critical reflection. In this sense, AI tools function as an extension of the teacher's professional practice rather than as its substitute.

ChatGPT as a pedagogical mediator

ChatGPT represents one of the most widely adopted platforms based on generative artificial intelligence models, and it substantially influences the way teachers plan and shape instructional activities. In practice, it is most commonly used to support the preparation of explanations, examples, and tasks, which enables faster lesson preparation and facilitates the differentiation of content according to students' knowledge levels and learning needs. Recent studies indicate that ChatGPT can serve as a "cognitive partner" in the instructional planning process, particularly in phases of conceptualizing lesson ideas, generating examples, and designing learning scenarios [9].

However, the use of ChatGPT requires critical and pedagogically informed oversight from the teacher. Holmes, Bialik, and Fadel [12] emphasize that generative models produce linguistically convincing text that may nonetheless be factually inaccurate, imprecise, or conceptually oversimplified. For this reason, pedagogical mediation remains essential—the teacher does not accept the generated content as final, but instead evaluates, selects, adapts, and interprets it in accordance with curricular standards and instructional objectives. Thus, the use of ChatGPT in educational settings is not a technical issue, but a didactic one.

Furthermore, empirical research highlights the risks associated with excessive or uncritical reliance on dialogic AI systems. Zhai, Wibowo, and Li [11] argue that frequent use of ChatGPT may reduce students' cognitive autonomy, as learners may replace the process of understanding with the reproduction of AI-generated responses. Similarly, the Pew Research Center reports that many teachers express concern that the use of ChatGPT could weaken the development of critical thinking and independent problem-solving skills if clear pedagogical guidelines are not established [6].

In other words, ChatGPT is only as effective as the teacher's competence to guide its use. When teachers possess well-developed digital and AI literacy, ChatGPT can expand instructional creativity, inclusiveness, and differentiation. When such competencies are lacking, there is a risk that the tool may diminish learning quality. This once again affirms that AI does not replace the teacher—it underscores the significance of the teacher's professional role.

Copilot as an integrated digital teaching partner

Microsoft Copilot, integrated within the Microsoft 365 platform, functions as an AI assistant embedded in applications already familiar to teachers, such as Word, PowerPoint, and Teams. This makes Copilot particularly relevant in educational environments where schools are already organizationally and technically aligned with the Microsoft ecosystem. In practice, Copilot is most commonly used to reorganize, expand, or condense existing instructional materials, as well as to generate presentation structures, lesson outlines, and class plans. In Tan's [9] research, Copilot was identified as a tool that reduces teachers' cognitive load and time burden during the lesson planning phase, allowing them to devote more attention to pedagogical analysis and instructional decision-making.

However, Copilot does not independently generate pedagogically meaningful content. Research on large language model interaction demonstrates that the quality of output depends directly on the quality of input—that is, on the clarity of goals, context, and instructions provided by the teacher [12]. If the teacher does not define a clear instructional purpose or assessment criteria, Copilot cannot ensure a meaningful educational solution. Its effectiveness is therefore highest when used as a tool for structuring and guiding the teacher's work, rather than as a source of original content. In this way, Copilot reinforces rather than replaces the teacher's expertise and planning role.

Grammarly and metacognitive support in writing

Grammarly is a tool for automated language analysis and correction that is increasingly used in the teaching of both native and foreign languages, as well as in subjects where written expression is of particular importance. The tool enables students and teachers to identify errors in sentence structure, spelling, syntax, and style. Elsayed et al. [2] demonstrate that the use of tools such as Grammarly can reduce writing-related anxiety and enhance students' motivation to write, especially when it is used as a form of reflective feedback.

However, the primary pedagogical value of Grammarly does not lie in its ability to "correct" a text, but in its potential to support learning about one's own writing. If students use the tool without understanding the reasons behind the corrections, there is a risk that they will not develop autonomy in writing, and that their individual style will become generic. For this reason, it is recommended that Grammarly be used after the initial writing process, as a means of reflection rather than during the drafting stage [12]. By doing so, students learn to recognize patterns in their own errors, which contributes to the development of linguistic awareness and metacognitive control in writing.

Perplexity AI as a tool for developing information literacy and critical thinking

Perplexity AI differs from most dialogic AI systems in that it provides the sources on which its responses are based. In an educational context, this feature carries strong pedagogical value, as it enables students and teachers to verify the credibility of information, compare sources, and develop a habit of critically examining knowledge. In studies focused on student-led inquiry and research, Perplexity has been shown to function as a valuable bridge between independent exploration and digital assistance [3].

Unlike ChatGPT, which may produce text that is linguistically convincing but not necessarily accurate, Perplexity promotes a culture of source validation, aligning with the principles of information literacy. Its use is particularly meaningful in secondary and higher education, as well as in project-based learning, where students are expected to develop the ability to connect, interpret, and evaluate information rather than merely reproduce content.

Canva AI and the visual articulation of learning content

Canva AI enables the generation of visual representations, diagrams, graphic posters, and presentations with minimal technical skill, which makes it accessible to teachers across different subject areas. Visualizing complex concepts can support comprehension for students who prefer visual learning styles and can foster creative expression in the classroom. However, as Holmes and Tuomi [13] caution, visually appealing design does not necessarily lead to deeper understanding. If Canva's visual output is used before the content has been meaningfully internalized, there is a risk that students may remain at a surface level of learning.

For this reason, Canva AI has the greatest pedagogical value when used in the later stages of learning, as a form of summarization, conceptual mapping, and knowledge presentation, rather than as an initial source of explanation. In this way, visual representation becomes evidence of understanding rather than a substitute for it.

Methods and materials

The primary research instrument was a purpose-designed online questionnaire distributed to teachers of different subject areas employed in one primary and one secondary vocational school in the Republic of Serbia. The questionnaire consisted of 15 items and covered the following areas: demographic information, level of digital literacy, frequency of AI tool usage, specific instructional contexts in which AI assistants are applied, perceived barriers, and attitudes regarding the ethical dimensions of artificial intelligence use in education. The items included multiple-choice questions, Likert-scale statements, linear rating scales, and open-ended questions for qualitative insights.

The sample consisted of 109 teachers, and data analysis was conducted using Excel and Python. Descriptive statistics were applied to present the frequencies and percentage distributions of responses, while hypothesis testing was performed using correlational analysis (Spearman's correlation). Open-ended responses were qualitatively analyzed through the identification of thematic patterns.

The main goal of the study was to examine how teachers perceive the role and potential of AI assistants in education, which AI tools they use, in which instructional situations, which barriers they identify, and which ethical dilemmas they consider important for the safe and responsible integration of artificial intelligence into teaching practice. The research was guided by the following hypotheses:

H1: Most teachers use AI assistants occasionally or experimentally, but not systematically in the instructional process.

H2: There is a positive correlation between the level of digital literacy and the frequency of AI tool use in education.

H3: The main barriers to broader use of AI assistants relate to insufficient knowledge, ethical concerns, and fear of misuse.

H4: Teachers who use AI tools more frequently express more positive attitudes toward their role in improving instruction and student motivation [1].

Results

The sample in this study is predominantly female, with 78.9% of respondents identifying as women and 21.1% as men. The age structure indicates a strong presence of more experienced teachers: 43.1% are over 51 years old, while 42.2% are between 41 and 50. Younger teachers are less represented, with 9.2% aged 31–40 and only 5.5% under the age of 30. Regarding educational attainment, most respondents hold a higher education degree—73.2% have completed undergraduate studies, 26.6% hold a master's degree, and 1.1% hold a doctoral degree. In terms of subject areas, the largest proportion of teachers work in the natural sciences (34.9%) and vocational subjects (26.6%), followed by social sciences (17.4%), informatics (15.6%), while arts teachers are the least represented (5.5%).

A large majority of teachers (86.2%) report using artificial intelligence in everyday life, while 11.9% do not, and 1.8% are unsure. However, when assessing their level of knowledge about AI, most respondents position themselves in the middle range (39.4%). A further 31.2% consider their knowledge to be low, and 16.5% very low. Only 11% report a high level of knowledge, and 1.8% describe their knowledge as very high. Regarding the use of AI assistants in teaching, the majority state that they currently do not use AI assistants but plan to in the future (45.9%). Occasional use is reported by 31.2%, regular use by 8.3%, while 14.7% neither use AI assistants nor plan to do so. Concerning familiarity with ethical principles related to the use of AI in education, 51.4% of respondents state that they are not familiar with them, 34.9% are partially familiar, and only 13.8% fully familiar.

When it comes to attitudes toward the impact of AI tools in education, most teachers express moderate optimism. They agree that AI assistants can contribute to improving instructional quality, yet they simultaneously express concern about potential misuse. Many respondents perceive AI as a time-saving resource, but emphasize the need for clear legal regulations and responsible use by students. The emotional stance toward the introduction of AI in education is generally positive or marked by curiosity (more than

half of respondents), though accompanied by a degree of caution [1].

Testing Hypothesis H1: Most teachers use AI assistants occasionally or experimentally, but not systematically in the instructional process

The results clearly confirm this hypothesis. The largest proportion of respondents (45.9%) reported that they do not currently use AI assistants in teaching, but intend to do so in the future. Occasional use was reported by 31.2% of teachers, while only 8.3% indicated regular use. A total of 14.7% of respondents stated that they do not use AI assistants and do not plan to introduce them.

This distribution indicates that the use of AI in education is presently situated in a phase of exploratory and preliminary adoption, where teachers experiment with these tools rather than implement them systematically. This suggests that AI assistants are not yet integrated into instructional practice as stable pedagogical resources, but rather appear as optional or supplementary tools [1].

Figure 1 presents the distribution of AI assistant usage among teachers, showing that occasional use and anticipated future use are notably more common than regular or systematic use.

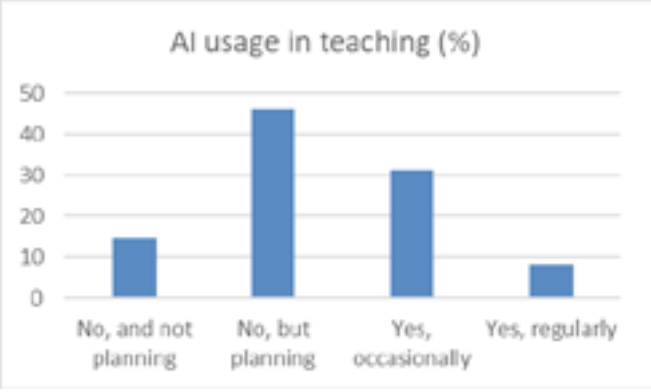


Figure 1. *Frequency of AI Assistant Use in Teaching*

Testing Hypothesis H2: There is a positive correlation between the level of digital literacy and the frequency of AI tool use in education

The Spearman correlation between self-assessed knowledge of AI technologies and the frequency of AI assistant use in teaching is $\rho = 0.36$, indicating a moderately positive relationship. This result supports the hypothesis that a higher level of digital literacy encourages more frequent use of artificial intelligence in the educational context. Teachers who are more confident in working with AI tools are more likely to incorporate them into lesson planning and instructional activities, even if only occasionally [1].

Testing Hypothesis H3: The main barriers to the broader use of AI assistants relate to insufficient knowledge, ethical concerns, and fear of misuse

The qualitative analysis of open-ended responses regarding the barriers to using AI in teaching shows that the most frequently mentioned concerns relate to misuse (38 occurrences), followed by fear (28 occurrences), and lack of knowledge or understanding (11 occurrences). These results indicate that teachers' most prominent concerns are associated with the potential for unethical use and manipulation of AI-generated content, as well as feelings of uncertainty stemming from insufficient training and limited understanding of how these tools function. In contrast, explicitly articulated ethical terms such as "ethics" or "safety" appeared less frequently in the responses, which may suggest that ethical dilemmas are experienced more as emotional apprehension than as clearly formulated professional concepts. These findings strongly confirm the hypothesis that insufficient knowledge, ethical ambiguity, and fear of misuse represent the key barriers to broader integration of AI assistants in educational practice [6].

Testing Hypothesis H4: Teachers who use AI tools more frequently have a more positive attitude toward their role in improving instruction and student motivation

The analysis of the relationship between the frequency of AI assistant use and attitudes measured on the Likert scale revealed several positive, although differently expressed, correlations. The strongest association was observed for the statement that "AI assistants save teachers' time" ($\rho = 0.40$), suggesting that those who use these tools more frequently recognize their practical value in everyday instructional work. A more moderate positive correlation was found for the statement that AI can contribute to improving the quality of teaching ($\rho = 0.21$).

At the same time, a slight negative correlation was observed in relation to concerns about misuse and the need for regulation, indicating that more frequent users of AI tools tend to express somewhat lower levels of fear regarding potential risks. These findings confirm that practical experience contributes to more positive attitudes and greater confidence in the usefulness of AI technologies in education [6].

DISCUSSION

The results of this study indicate that the use of AI assistants in education among teachers in Serbia is still at an early stage of integration and is occurring primarily on an individual rather than a systemic level. Most respondents do not use AI tools regularly but perceive them as a potential form of support that could be incorporated into their professional practice in the future. Similar patterns have been observed in international research, where teachers recognize the potential benefits of AI in education but remain uncertain about how to apply such tools in a pedagogically meaningful and responsible way. A key factor underlying this caution is the lack of formal training and clear implementation guidelines.

The observed moderate positive correlation between teachers' digital literacy and the frequency of AI tool use suggests that higher levels of digital self-efficacy directly influence openness to technological innovation. Teachers with more developed digital competencies are better prepared to evaluate the pedagogical value, limitations, and practical implications of AI tools, and are therefore more likely to experiment with and integrate them into instruction. This finding supports the view that digital competence represents an important prerequisite for the responsible use of AI in educational contexts.

Although teachers' attitudes toward AI are generally positive—particularly regarding the potential for saving time and facilitating the preparation and adaptation of instructional materials—significant concerns remain. The most frequently identified risks relate to the possibility of misuse, challenges to academic integrity, and the danger that students may become overly dependent on AI tools instead of developing independent thinking and problem-solving skills. These concerns reflect broader debates in the literature, which warn that uncritical application of conversational AI systems can lead to reduced cognitive autonomy among learners.

Importantly, both the findings of this study and existing research consistently emphasize the continued central role of the teacher, regardless of the capabilities of AI systems. The most favorable learning outcomes are reported when AI support is combined with expert pedagogical guidance. While AI can enhance the efficiency and organization of instruction, it cannot replace the teacher's professional judgment, emotional interaction, and reflective decision-making, which remain essential in the learning process.

Finally, the results show that teachers who use AI tools more frequently tend to hold more positive attitudes toward them, while those with less experience are more likely to express hesitation or concern. This suggests that practical experience contributes to developing confidence and the ability to use AI critically, intentionally, and in a pedagogically responsible manner.

RECOMMENDATIONS FOR EDUCATIONAL PRACTICE AND TEACHER COMPETENCE DEVELOPMENT

The results of the study indicate that the successful integration of artificial intelligence into the educational process requires coordinated effort at both the level of classroom practice and the education system as a whole. AI tools can enhance instructional efficiency, support the adaptation of learning materials to students with different abilities, and facilitate lesson preparation, but only when they are used within clearly defined pedagogical objectives and when teachers possess adequate digital and AI literacy. Therefore, it is essential to ensure professional development that focuses not only on understanding the functional capabilities of AI tools but also on developing a critical stance toward them, particularly with regard to evaluating the quality of generated information, ensuring student data protection, and maintaining academic integrity.

At the level of school practice, it is recommended to introduce pedagogical scenarios that clearly specify when and how AI can be integrated into the teaching process. These scenarios should reflect the different phases of learning: during the initial acquisition of concepts, AI may serve as visual or explanatory support, whereas in phases that involve analysis,

argumentation, and reflection, the use of technology should be limited in order to preserve students' cognitive autonomy. It is especially important to encourage students to explain, rephrase, and critically evaluate AI-generated outputs rather than accept them passively.

From a systemic perspective, it is recommended to develop continuous professional development programs in which teachers are trained not only to use AI tools technically but also to apply them responsibly and inclusively in pedagogical practice. Such training should integrate three key dimensions: understanding the algorithmic process and limitations of AI, pedagogical-didactic use of AI in lesson planning and implementation, and ethical principles and data protection standards. Furthermore, it is crucial to provide clearly formulated guidelines for the responsible use of AI, expressed in the form of accessible and concrete recommendations that are understandable to teachers with varying levels of digital competence.

Framework for the responsible integration of AI in teaching

The introduction of artificial intelligence tools into the educational process cannot be viewed as a technical matter alone, but rather as a pedagogical and ethical decision made by the teacher. The effectiveness and safety of AI use in teaching depend on how the teacher plans, guides, monitors, and reflects on its application. In this sense, artificial intelligence does not replace the teacher; instead, it functions as an instrument that can support the teacher's professional practice when used responsibly and critically.

The integration of AI in teaching requires a deliberate and guided process in which the teacher clearly defines the learning objectives, the context of use, and the scope of activities in which AI will be included. The teacher's key role lies in understanding when AI use is meaningful and when it may undermine students' independent thinking, creativity, or academic integrity. In other words, AI may facilitate learning, but it must not take over the learning.

The following section presents a model for the responsible integration of AI tools into the teaching process, based on the didactic principles of planning, mediation, and reflection:

1. Pedagogical formulation of goals and purpose

The teacher first defines what the student is expected to learn and only then decides whether and how AI can support that process. If the lesson objective involves acquiring new knowledge, AI may be used as a medium for explanation and illustration. However, if the objective requires independent analysis, evaluation, or creativity, AI use must be limited to supportive functions rather than generating final answers. This helps protect the development of critical thinking and cognitive autonomy.

2. Pedagogical control and verification of content

AI-generated content may be linguistically convincing but incomplete or factually inaccurate, as noted by Holmes, Bialik, and Fadel [12]. Therefore, the teacher assumes the role of source validation and quality assurance. The teacher verifies, adjusts, and aligns content with curricular standards and the classroom context. AI may provide material, but the teacher gives it meaning.

3. Active role of the student

The student should not become a passive recipient of AI-generated answers. Learning occurs through effort, error, questioning, and understanding. AI should therefore be used as a partner in dialogue, not as an authority. Students should be encouraged to explain, reformulate, challenge, or expand AI-generated content. This approach fosters metacognition, autonomy, and reflection, as suggested by studies such as Elsayed [2]

4. Ethics, transparency, and academic integrity

The use of AI tools in education requires clearly defined boundaries. Teachers should openly communicate to students when AI use is permitted, which tasks must be completed independently, how to verify the reliability of information, and why plagiarism — including "AI-assisted plagiarism" — undermines the development of essential learning skills.

Conclusion

The results of the study confirm that artificial intelligence can serve as a valuable resource for enhancing the educational process, but only when its use is guided by well-considered pedagogical intentions and responsible implementation practices. Although teachers recognize the benefits of AI assistants in saving time, preparing instructional materials, and differentiating content, the adoption of these tools is not yet systemically supported and instead depends largely on individual initiative, personal experience,

and the level of digital competence. This situation indicates the need for organized support for teachers, through structured training programs and guidelines that integrate the technical, pedagogical, and ethical dimensions of AI use in teaching.

The teacher's role remains central, as learning does not consist merely of reproducing information, but involves effort, error, questioning, and understanding. AI tools can contribute to learning only when they are used as partners in dialogue, rather than as authorities or substitutes for students' own cognitive processes. Students should be encouraged to explain, challenge, reformulate, and critically evaluate AI-generated content, thereby fostering metacognition, autonomy, and reflective thinking, as emphasized in recent research. For this reason, clearly distinguishing the role of the teacher from the role of AI remains a key principle in the responsible use of educational technologies.

In conclusion, the successful integration of artificial intelligence in school practice does not depend solely on the technical availability of tools, but on the pedagogical competence of teachers and the existence of educational policies that support their professional autonomy. AI in education has the potential to contribute to more inclusive, flexible, and motivating learning environments, but only if it remains aligned with educational objectives and does not replace human interaction, professional judgment, or instructional responsibility. Therefore, the development of AI literacy among teachers represents a crucial step toward the responsible, ethical, and meaningful use of these technologies in education.

Acknowledgements

The author would like to express gratitude to the participating teachers for their time and cooperation, as well as to the school administration for supporting the implementation of this research. Special thanks are also extended to colleagues who provided feedback and guidance throughout the development of the study.

REFERENCES

[1] A. Ivanov, J. Aleksić, Z. Avramović, Ž. Stanković, L. Stošić, and O. Krčadinac, "AI asistenti kao podrška nastavnicima," in Proceedings of the ITeo Conference, Banja Luka, 2025.

[2] D. Elsayed et al., "AI-assisted instruction and learner outcomes in language education," Language Testing in Asia, 2024.

[3] I. M. García-L López and L. Trujillo-Liñán, "Ethical and regulatory challenges of generative AI in education: A systematic review," Frontiers in Education, vol. 10, June 2025.

[4] M. Khalifa and M. Albadawy, "Using artificial intelligence in academic writing and research: An essential productivity tool," Computer Methods and Programs in Biomedicine Update, vol. 5, p. 100145, 2024.

[5] L. Labadze et al., "A systematic review of AI chatbots in education," International Journal of Educational Technology in Higher Education, 2023.

[6] L. Lin, "A quarter of US teachers say AI tools do more harm than good in K–12 education," Pew Research Center, 2024.

[7] M. H. Ling, "Artificial intelligence in teaching and teacher professional development: A systematic review," Computers and Education: Artificial Intelligence, vol. 8, p. 100355, 2025.

[8] L. Stošić, A. Radonjić, and O. Krčadinac, "The evolution of artificial intelligence and its transformative role in education," Journal of UUNT: Informatics and Computer Sciences, vol. 2, no. 1, pp. 15–21, 2025. doi: 10.62907/juuntics250201015s

[9] X. Tan, "Teachers' integration of AI in instructional design: Needs and gaps in training," Education and AI Journal, 2024.

[10] X. Yan et al., "LLM in education: Risks and ethics," arXiv preprint, 2023. Available: https://arxiv.org/abs/2303.13379

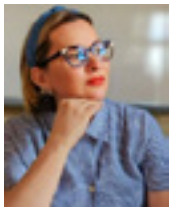
[11] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review," Smart Learning Environments, vol. 11, no. 1, p. 28, 2024.

[12] W. Holmes, M. Bialik, and C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Boston, MA: Center for Curriculum Redesign, 2019.

[13] I. Tuomi, The Impact of Artificial Intelligence on Learning, Teaching, and Education. Luxembourg: Publications Office of the European Union, 2018.

Received: November 4, 2025
Accepted: November 18, 2025

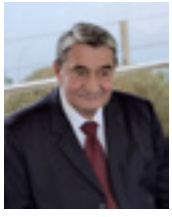
About the authors



Aleksandra Ivanov was born on September 19, 1979, in Belgrade. After completing high school, she enrolled in undergraduate studies at the Faculty of Organizational Sciences in Belgrade. With an excellent grade on the defense of her thesis titled "Enterprise Restructuring," she obtained the title of Graduate Engineer of Organizational Sciences. In 2012, she enrolled in master studies at the Faculty of Technical Sciences in Čačak, and with the highest grade on the defense of her master's thesis "ICT Literacy of Teachers at Dragojlo Dudić Primary School," she earned the degree of Master Professor of Technical and Computer Education (2015). In 2022, she began her doctoral studies at the Faculty of Informatics and Computing, Union Nikola Tesla University, in the study program Informatics.

The education of students represents the core of her professional career. From 2006 to 2007, she worked as a mathematics teacher at Duško Radović Primary School. She then spent ten years as a teacher of Informatics and Computing at Dragojlo Dudić Primary School (2007–2017). After that, she joined the Architectural Technical School in Belgrade as a professor of Informatics and Computing, where she continues to work to this day.

She speaks Serbian, English, and Bulgarian. She is a co-author of four textbooks, reviewer of two textbooks, and facilitator of three accredited professional development seminars. She has presented her research at numerous conferences and has published papers in several academic journals.



Academician, professor emeritus, ex-rector of the University **Zoran Ž. Avramović**, PhD, was born in Serbia. He finished elementary and high school with great success. He was awarded several diplomas by Nikola Tesla and Mihailo Petrović Alas. He graduated on time at the University of Belgrade - Faculty of Electrical Engineering, with an average grade of 9.72 in five-year studies. He received his master's degree at that faculty (all excellent grades, exams and master's degrees), and then obtained a doctorate in technical sciences (in 1988). As an excellent student of the University, he had the right and at the same time studied mathematics at the Faculty of Mathematics in Belgrade. He was the champion of Serbia in mathematics ("first prize") and Yugoslavia in electrical engineering ("gold medal"). Winner of the Gold Medal of the Russian Academy of Sciences for Merit in Electrical Engineering. He was professor of the Leading Russian National Research University – HSE.

Olja Krcadinac (Latinovic, maiden name) is assistant professor at "Union – Nikola Tesla" University - Faculty of Informatics and Computer Science. She earned her Ph.D. in biometric field from University of Belgrade – Faculty of Organizational science, where she conducted groundbreaking research on speaker recognition. In addition to her teaching responsibilities, Olja has authored numerous impactful publications in peer-reviewed journals, contributing valuable insights to the scientific community. Her research focuses on biometric, sensors, IoT and AI, addressing critical issues in AI and making significant contributions to the academic community.

Zeljko Stankovic received his higher education in Cleveland, Ohio, USA, where he graduated in 1981. The topic of the thesis was "Reversible sound in halls". He defended his master's thesis ("Learning control system (LMS) based on ADL SCORM specifications") in 2006 at the University of Novi Sad, Faculty of Science, Department of Informatics. He defended his doctoral dissertation (Laser perception of defined objects and encapsulation of control and logic elements for an autonomous robotic teaching tool) at Singidunum University, Belgrade, in 2010. He has been programming since 1984, creating programs for his first Commodore 64 computer. He works as a full-time professor at Pan-European University "APEIRON". Robotics and bioengineering have been a field of work and interest for many years. He is the holder of the patent right for the teaching tool CD ROBI.

For citation

Aleksandra Ivanov, Zoran Ž. Avramović, Olja Krčadinac, Željko Stanković, The Role of AI Assistants in Supporting Teachers, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:113-121, (UDC: 371.13:004.738.5), (DOI: 10.7251/JIT2502113I), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

AI-DRIVEN TRANSFORMATION OF THE FITNESS INDUSTRY: A CASE STUDY OF G&S PREMIUM GYM

Vesna Radojcic[1], Milos Dobrojevic[2]
[1] *University Sinergija, Faculty of Computing and Informatics, Bijeljina, Bosnia and Herzegovina, vradojcic@sinergija.edu.ba, 0000-0001-7826-1081*
[2] *University Sinergija, Faculty of Computing and Informatics, Bijeljina, Bosnia and Herzegovina, mdobrojevic@sinergija.edu.ba, 0000-0003-3798-312X*

Original scientific paper
https://doi.org/10.7251/JIT2502122R UDC: 796.01/.09:004.021

Abstract: This paper explores the transformative impact of digital technologies on the fitness industry, focusing specifically on the role of artificial intelligence (AI) in enhancing modern exercise equipment. By analyzing the needs of recreational users and professional athletes, it examines how AI-driven fitness devices optimize personalization, improve performance tracking, and elevate the overall user experience. Using G&S Premium Gym in Bijeljina—the first fitness center in Bosnia and Herzegovina to integrate AI-powered equipment—as a case study, the paper delves into the technical specifications, functionality, and user interaction with these advanced machines. Key findings reveal that AI technologies significantly enhance training efficiency and customization, contributing to measurable improvements in user satisfaction and physical performance. However, challenges persist, particularly regarding technology accessibility, user digital literacy, and data privacy concerns. The research highlights the potential for AI to redefine standards in recreation while addressing these challenges. Recommendations for future research and implementation emphasize the importance of affordable, user-friendly AI solutions and improved data security measures.

Keywords: Artificial intelligence (AI), digital transformation, fitness industry, smart exercise equipment, training personalization

INTRODUCTION

Modern society is undergoing rapid technological advancements that profoundly impact health and physical fitness practices. AI-powered systems are revolutionizing gym management by optimizing operational efficiency and automating tasks, while also enabling hyper-personalized workout and nutrition plans through adaptive algorithms [1] [2]. Global leaders such as Technogym have spearheaded these innovations by introducing solutions that enhance training personalization, precise performance tracking, and increased user motivation. The smart fitness equipment market, valued at several billion dollars, is projected to grow annually by over 20%, highlighting the increasing demand for these technologies [3].

G&S Premium Gym in Bijeljina, equipped with AI-powered devices, exemplifies how digital transformation reshapes the recreation sector by setting new standards in training personalization, user interaction, and data analytics. This paper examines the impact of AI-driven modern exercise equipment on training paradigms, user experience, and technological perceptions in the fitness industry. Using G&S Premium Gym as a case study, it investigates the technical characteristics and functionalities of smart equipment while addressing associated challenges.

THEORETICAL FRAMEWORK

Artificial Intelligence in Modern Industries

Artificial intelligence (AI) is driving innovation across many industries by transforming traditional processes and setting new operational standards. In healthcare, AI aids diagnostics, treatment planning, and patient monitoring, improving early disease detection and outcomes [4]. In education, AI platforms personalize learning by adapting to individual student needs. [5].

Manufacturing benefits from intelligent robotics and automated quality control, boosting efficiency

and reducing errors [6]. Retail uses AI for recommendation systems, chatbots, and supply chain optimization [7]. The fitness industry is increasingly adopting AI through smart equipment, virtual trainers, and wearables, enhancing physical activity with data-driven, personalized experiences.

Research shows AI optimizes training, boosts engagement, and monitors fitness results, though challenges remain in accessibility and adaptability for diverse users. Despite these advances, AI adoption raises concerns around ethics, data privacy, and digital literacy [8]. Balancing innovation with responsibility is key to successful AI integration across sectors.

Digital Transformation in the Field of Sports and Recreation

The ongoing digital transformation in the fitness sector is changing how people exercise and manage health. The integration of artificial intelligence (AI) and Internet of Things (IoT) technologies enables more personalized fitness experiences through virtual assistants, smart wearables, and advanced equipment [9], [10]. These tools offer flexible workouts, track vital health data, and provide analytics-based insights to improve performance.

For example, AI-generated calisthenics programs effectively improve specific fitness metrics, although human-designed programs may still offer greater adaptability and nuance [11]. AI-driven solutions also moderately boost short-term activity (like step counts), but long-term behavior change requires enhanced human-AI collaboration [12]. Combining AI with social-IoT frameworks fosters the sharing of user experiences and better fitness outcomes. While these innovations help combat sedentary lifestyles in a digital world, challenges remain, particularly balancing automation with human expertise to maintain engagement and address individual differences [13]. As the industry evolves, striking this balance will be crucial for promoting global physical activity.

Research Methodology

This research used a mixed-methods approach, combining qualitative and quantitative techniques to thoroughly analyze the use of AI technologies at G&S Premium Gym. Data collection methods included:

- **Interviews:** Conducted with key stakeholders—trainers (Nt = 3) and gym members (Nu = 9)—to gain insights into the practical use and impact of AI-powered smart equipment. To ensure objectivity and reproducibility, the following questions were asked during interviews:
  1. *How do you evaluate the use of AI equipment compared to traditional machines?*
  2. *Do AI features (e.g., adaptive load, real-time feedback, personalized plans) influence your motivation?*
  3. *What are the main advantages and disadvantages of AI equipment?*
  4. *Would you recommend AI equipment to other users?*
  5. *How easy is it to learn to use AI functionalities?*
- **Technical Documentation Analysis:** Reviewed detailed specifications and operational reports from AI-integrated devices like Technogym Biostrength equipment and smart scales.
- **Equipment Testing:** Evaluated smart devices during regular gym sessions to assess features such as adaptive load, motion tracking, and performance analytics.
- **Feedback Analysis:** Collected and categorized user feedback to identify trends related to motivation, usability, and technical issues.

Though the sample size was relatively small, it provides a solid foundation for exploring AI's practical application and impact in fitness. Insights from trainers and gym members offer valuable initial findings to inform future, larger studies. Ethical standards were strictly followed; all participants gave informed consent, and their data were anonymized and used solely for research purposes, ensuring privacy and legal compliance. Participants' ages ranged from 15 to 45 years. This methodology enabled a robust and detailed examination of how AI-powered equipment affects user experience, engagement, and fitness outcomes, forming a strong basis for the study's findings and discussion.

Case Study: G&S Premium Gym Bijeljina

G&S Premium Gym, located in Bijeljina, covers 2,400 square meters, making it one of the largest and most modern fitness centers in the region. It offers diverse training programs for both recreational users and professional athletes, aiming to improve health and fitness through advanced technology and

personalized training. As the first gym in Bosnia and Herzegovina to implement AI-powered smart equipment, G&S Premium Gym provides a unique training experience. Figure 1 shows an example of the AI-enabled equipment used at the center.

Figure 1 - *AI-Powered Smart Equipment at G&S Premium Gym*

Smart Equipment with AI Integration

G&S Premium Gym is a pioneer in using smart equipment powered by AI to optimize workouts and track user progress. The gym's equipment is developed by Technogym, whose ecosystem includes interconnected devices like treadmills, strength machines, and smart scales that collect and analyze biometric data in real time with integrated AI algorithms. These devices enable personalized exercise by measuring key body parameters and analyzing performance. Key Features of Smart Equipment:

- **Adaptive Load:** Machines automatically adjust weight or intensity based on users' abilities and goals.
- **Motion Tracking:** Advanced technology precisely analyzes movements to ensure correct exercise execution.
- **Performance Analytics:** Devices generate detailed reports on calories burned, strength levels, and progress toward goals.

Examples include smart treadmills that monitor heart rate and pace, ergometers with motivational and training screens, and smart scales measuring body composition. These machines use AI and patented aerospace technology to adapt resistance according to the user's neuromuscular profile. The Biostrength system adjusts load and tempo in real time, provides movement guidance, motivational feedback, and customizes programs based on user goals and progress.

Unlike traditional strength machines, Biostrength precisely controls both concentric and eccentric phases, enabling scientifically optimized workouts and lowering injury risk [14]. This system benefits beginners and professional athletes alike by improving strength, muscle balance, and performance through highly personalized protocols [15].

This AI-powered interface, illustrated in Figure 2, demonstrates how Technogym's Biostrength equipment delivers real-time feedback during strength training sessions. The system automatically adjusts workload across sets, tracks the number of completed repetitions, and provides visual guidance for each exercise. Additionally, it employs load progression algorithms to optimize muscular adaptation, ensuring both safe and efficient strength development.

Figure 2 - Biostrength Equipment Interface: Real-Time Load Adjustment and Rep Tracking

Bluetooth-Connected Equipment Integration

The integration of Bluetooth-connected equipment expands the app's capabilities, providing seamless connectivity with smart fitness devices. Users can pair their devices, such as smart scales, treadmills, or ergometers, directly within the app to access real-time performance data and analytics. Features of the Bluetooth Integration:

- Device Pairing,
- Real-Time Tracking,
- Workout History Sync.

Figure 3 shows an example of the application interface.

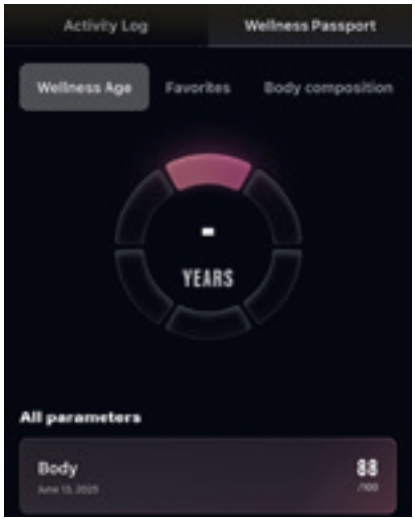


Figure 3 - Application Interface Overview

AI for beginners: Facilitating the First Steps in the Gym

Artificial intelligence is especially valuable for gym beginners who struggle with equipment use, weight selection, and program design. AI addresses these challenges by offering automated recommendations, personalized plans, and guided workouts. For instance, adaptive load equipment adjusts weights based on users' initial measurements. Visual and audio instructions help reduce injury risk, while progress tracking and gamification boost motivation. Structured training circuits designed for beginners enable effective targeting of key muscle groups, helping even novices achieve their goals confidently and safely.

User Experience and Feedback

AI provides precise feedback and visual analytics to help users reach their goals. For example, a smart scale measures body weight, fat percentage, muscle mass, and hydration, syncing data automatically with an app for continuous monitoring and personalized advice [16]. These innovations simplify tracking and boost motivation through interactive features and gamification (Figure 4). Trainers note that AI equipment eases program personalization and progress tracking, though users sometimes need extra help to learn its functions. Continuous feedback collection is crucial to improve features and adapt to different skill levels. This equipment supports more efficient training planning, enhancing user satisfaction and fitness center organization.

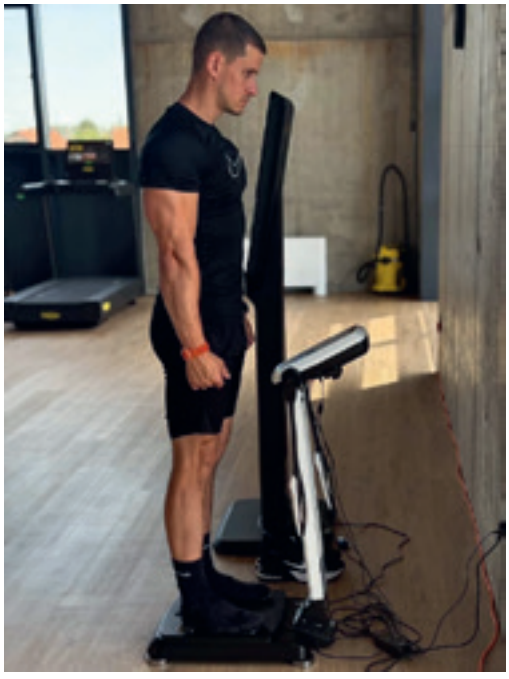


Figure 4 - Smart Scale in Action: Personalized Fitness Insights

Figure 5 - User Interface of the Smart Scale: Real-Time Data and Analytics

Figure 5 shows how the smart scale and its software display real-time measurement results. The key parameters measured are:

- Muscle mass: 44.4 kg, rated as "GOOD"
- Total body water: 57.5%, rated as "HIGH"
- Body fat: 22.5 kg or 20.7%, rated as "AVERAGE"
- Basal metabolic rate: 1907 kcal
- Body weight: 91.8 kg

These are measured using bioelectrical impedance with AI algorithms built into the Technogym device. The software automatically categorizes the results like "GOOD," "HIGH," or "AVERAGE," taking into account the user's gender, age, height, and other info.

The results are analyzed in several ways:

- Calculating averages, minimums, and maximums in the group
- Comparing measurements before and after training to track progress
- Showing changes visually over time with graphs
- Exploring relationships between parameters, like how hydration affects performance

Professional trainers interpret the results with help from the system, which also gives personalized advice on diet and training so users get recommendations tailored just for them.

DISCUSSION

The study underscores the transformative potential of artificial intelligence (AI) in fitness, particularly in enhancing training personalization and user engagement. The G&S Premium Gym case demonstrates how AI-powered equipment improves workout efficiency and elevates the user experience with real-time feedback and personalized insights. These findings align with broader trends in industries like healthcare and education, emphasizing efficiency, customization, and engagement. AI's ability to tailor workouts based on individual needs, through adaptive load functionalities and motion tracking, ensures alignment with user goals and reduces injury risks by encouraging proper techniques. Gamification elements in smart devices effectively maintain motivation, promoting long-term physical activity and healthier lifestyles. Despite these advantages, challenges persist, including the high cost of AI equipment, limited digital literacy, and critical concerns about data privacy [17]. Addressing these barriers requires cost-effective solutions, improved digital literacy, and robust data protection frameworks. Future research should explore these areas to ensure wider accessibility and responsible technology use. The integration of AI in fitness centers like G&S Premium Gym provides a blueprint for the future, showcasing how innovation can revolutionize training practices while emphasizing the need to overcome associated challenges.

Similar AI implementations already exist in premium fitness centers worldwide, such as Technogym's Biostrength system in Italy and the UAE, as well as home-based solutions like Peloton and Tonal in the United States. Additionally, applications such as Fitbod leverage AI to create personalized workout programs based on user data. These examples confirm the global trend of digital transformation in the fitness industry and highlight the growing importance of AI-driven personalization across different contexts.

CONCLUSION

The case of G&S Premium Gym illustrates how AI-powered equipment increases workout efficiency, ensures safety, and tracks progress with precision. Interviews with trainers (Nt = 3) and gym members (Nu = 9) reveal that trainers value AI for personalization and tracking, while users appreciate motivational data visualization, despite some initial technical issues, highlighting the need for user education.

Broader adoption of AI faces challenges like high costs, limited digital literacy, and data privacy concerns. Addressing these requires affordable technologies, robust data protection frameworks, and user education initiatives. Steps include gradual integration of AI, staff training, and collaboration with policymakers to balance innovation and privacy. Future research should explore cost-effective AI solutions for smaller gyms, demographic inclusivity, long-term health impacts of AI-driven programs, and advancements in predictive analytics. By overcoming challenges, AI can revolutionize fitness, making it more inclusive, efficient, and accessible to a wider audience.

Although certain AI-based solutions—such as predictive analytics, virtual coaching, and recovery monitoring—already exist in early forms, their full implementation in commercial fitness centers is still limited. Future developments are expected to enhance the accuracy of performance prediction, integrate wearables and smart equipment into unified ecosystems, and enable more advanced real-time adjustments during training.

REFERENCES

[1] K. C. Rohit, A. Mohd, M. Yadav, and S. Tripti, "The future of gym management: Harnessing the power of artificial intelligence," International Journal of Innovative Research in Computer Science & Technology, 2024. doi: 10.55524/csistw. 2024.12.1.50.

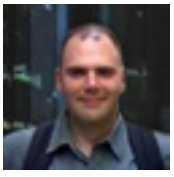
[2] K. Navale, "FitnessGPT using artificial intelligence and deep learning," International Journal for Research in Applied Science and Engineering Technology, 2024. doi: 10.22214/ijraset.2024.61477.

[3] D. Patel, "Smart fitness market," 2025. [Online]. Available: https://dataintelo.com/report/smart-fitness-market
[4] S. Research, "Eric Topol pens book on artificial intelligence in medicine," 2019. [Online]. Available: https://www.scripps.edu/news-and-events/press-room/2019/20190312-topol-deep-medicine.html
[5] L. Chen, P. Chen and Z. Lin, "Artificial intelligence in education: A review," IEEE, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9069875
[6] D. Ivanov, A. Dolgui, A. Das, and B. Sokolov, "Digital supply chain twins: Managing the ripple effect, resilience, and disruption risks by data-driven optimization, simulation, and visibility," Handbook of Ripple Effects in the Supply Chain, vol. 276, 2019. doi: 10.1007/978-3-030-14302-2_15.
[7] I. I. Brainvire, "AI & the future of retail: Personalization, automation, & predictive analytics," 2023. [Online]. Available: https://www.linkedin.com/pulse/ai-future-retail-personalization-automation-predictive/
[8] L. Floridi et al., "AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," Minds and Machines, 2018.
[9] B. Thomas, "Fitness technology: Virtual assistants, apps and apparatus in the cyberage," Studies in Indian Place Names, vol. 40, no. 3, 2020.
[10] F. Alireza, F. Reza, R. Javad, and M. Roberto, "Application of Internet of Things and artificial intelligence for smart fitness: A survey," Computer Networks, vol. 189, 2021. doi: 10.1016/j.comnet.2021.107859.
[11] R. C. E. Masagca, "The AI coach," Journal of Human Sport and Exercise, 2024. doi: 10.55860/13v7e679.
[12] G. E., L. Dillys, R.-R. Octavio and D. Kerstin, "Human factors in AI-driven digital solutions for increasing physical activity: Scoping review," JMIR Human Factors, 2024. doi: 10.2196/55964.
[13] Š. Valdemar et al., "Let's (Tik) talk about fitness trends," Frontiers in Public Health, vol. 10, 2022. doi: 10.3389/fpubh.2022.899949.
[14] M. Roig, K. O'Brien, G. Kirk, R. Murray, P. McKinnon, B. Shadga, and W. D. Reid, "The effects of eccentric versus concentric resistance training on muscle strength and mass in healthy adults: A systematic review with meta-analysis," British Journal of Sports Medicine, 2009.
[15] Technogym, "Biostrength – Technogym,". [Online]. Available: https://www.technogym.com/en-GB/
[16] C.-S. Giselle, R. Medically, and A. M. Nasha, "Best smart scale for 2025. Plus, expert tips on getting an accurate reading," 2025. [Online]. Available: https://www.cnet.com/health/fitness/best-smart-scale/
[17] D. Vioreanu, "Top AI trends shaping the fitness industry in 2025," 2025. [Online]. Available: https://3dlook.ai/content-hub/ai-in-fitness-industry/

Received: September 16, 2025
Accepted: November 18, 2025

About the authors

Vesna Radojcic is a Ph.D. student in Computer Systems Security at the University Sinergija and works as an assistant at the Faculty of Computing and Informatics. She holds a Master's degree in Contemporary Information Technologies, and her research interests include the application of artificial intelligence and computer vision technologies. She has participated in numerous scientific conferences.



Milos Dobrojevic, Ph.D., is an Associate Professor at the Faculty of Computing and Informatics, University Sinergija, Bijeljina, Republic of Srpska, Bosnia and Herzegovina. He earned his Ph.D. in 2006 from the Faculty of Mechanical Engineering, University of Belgrade, Republic of Serbia. He has published over 50 scientific papers, with a research focus on software engineering, artificial intelligence, and smart IoT applications.

For citation

Vesna Radojcic, Milos Dobrojevic, AI-Driven Transformation of the Fitness Industry: A Case Study of G&S Premium Gym, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:122-127, (UDC: 796.01/.09:004.021), (DOI: 10.7251/JIT2502122R), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

A Small Language AI Model in the Bosnian Language

Boško Jefić[1], Vlatko Bodul[2], Admir Agić[3]
[1] *the City Administration of Zenica, Zenica, BH, bosko.jefic@zenica.ba*
[2] *B.Sc.it, Zenica, BH, vlatko.bodul@unze.ba*
[3]*Kolektiv EU, Tešanj, BH, admir.agic@kolektiv.pro*

Original scientific paper
https://doi.org/10.7251/JIT2502128J UDC: 378.147.311.4:821.163.4

Abstract: This study presents the development and evaluation of Mali Mujo, a small-scale language model optimized for the Bosnian language, designed to operate efficiently on devices with limited computational resources. Leveraging the TinyLlama architecture, the model demonstrates the feasibility of deploying natural language processing (NLP) applications in environments with constrained memory and processing capabilities, specifically devices with 1 GB storage and 8 GB RAM. The system integrates Langchain agents and the DuckDuckGo API to enable real-time information retrieval, enhancing the model's responsiveness and accuracy in practical applications. The methodology involved training the TinyLlama model on a curated Bosnian dataset, followed by testing across diverse real-world scenarios in industry and administration. Performance metrics focused on accuracy, response time, and computational efficiency, while additional evaluation considered user experience and adaptability to domain-specific tasks. The results indicate that Mali Mujo delivers rapid and reliable responses to user queries, with significant advantages in speed and resource efficiency compared to larger language models. The model effectively processes administrative requests, generates technical and market-related insights, and supports educational and governmental applications, highlighting its versatility. While small-scale models exhibit lower absolute accuracy than their larger counterparts, the study demonstrates that careful optimization and integration with external APIs can mitigate limitations, providing a balance between performance and accessibility. Furthermore, the model's design ensures user privacy and low energy consumption, contributing to sustainable and secure AI deployment. Mali Mujo exemplifies the potential of small language models to enhance efficiency, accessibility, and usability in local-language contexts. Its deployment provides a scalable, cost-effective solution for organizations with limited infrastructure, offering opportunities for further enhancement through expanded datasets, multilingual support, adaptive learning, and integration with emerging AI technologies. The findings underscore the practicality of small AI models in bridging the gap between advanced NLP capabilities and resource-constrained environments.

Keywords: Small language models, TinyLlama, Bosnian language, Langchain agents, Real-time information retrieval, AI in industry

Introduction

The development of an AI model in the Bosnian language holds crucial importance for enhancing communication and efficiency within both industry and public administration. Although the Bosnian language is rich and diverse, it faces significant challenges in automation and data processing due to its specific grammatical and syntactic structures. Current AI models are often not optimized for operation on devices with limited computational resources, such as those with smaller storage capacities or lower amounts of RAM.

Implementing an AI model in the Bosnian language can facilitate administrative processes, enable faster documentation processing, and improve communication between citizens and institutions. By employing AI systems, particularly small language models, it is possible to automate many routine tasks, thereby saving considerable time and resources. Furthermore, such a model could enhance data organization, information retrieval, and the analysis of administrative documentation.

Developing a language model specifically for Bosnian ensures more accurate recognition of idiomatic expressions, terminology, and contextual nuances unique to the language.

In addition, this development plays a vital role in promoting digital inclusion by enabling a larger num-

ber of users to access modern technological solutions in their native language. Given that many individuals still rely on administrative services in their mother tongue, AI models can contribute to improved compliance with legal and regulatory frameworks.

Although AI models for other languages exist, the particularities of Bosnian require tailored approaches to ensure both accuracy and contextual relevance. Utilizing such a model allows for faster data recognition and processing, as well as greater reliability in decision-making processes. Moreover, AI models can help prevent errors in administrative workflows, which often depend on precise linguistic interpretation.

This opens up opportunities for improving user experience and increasing citizen satisfaction with public services. The development of these models also stimulates the digital transformation of sectors that continue to depend heavily on manual data entry. Ultimately, the deployment of AI models in the Bosnian language can strengthen Bosnia and Herzegovina's competitiveness on the global technological landscape [1].

Research Objective

The objective of this research is to develop a **small-scale AI language model** specifically adapted to the Bosnian language, capable of efficient operation on devices with limited hardware resources. In today's digital age, many users in Bosnia and Herzegovina, as well as across the broader region, rely on devices with modest technical specifications. This model is intended to be optimized for systems with as little as **1 GB of storage and 8 GB of RAM**, which presents challenges in terms of both speed and efficiency as defined in Table 1.

By developing such a model, the goal is to enable users who lack access to high-end technology to still benefit from advanced linguistic tools in their native language. In addition to being lightweight, the model must effectively handle the linguistic complexity of Bosnian, including its grammatical and syntactic features[2].

Table 1. *Key technical specifications and objectives of the Bosnian-language AI model*

Specification	Description
Target devices	Devices with 1 GB storage and 8 GB RAM
Model type	Small language AI model
Optimization	Designed for operation on resource-limited devices
Focus	Recognition and processing of the Bosnian language
Applications	Administrative and industrial processes
Linguistic complexity	Processing grammatical and syntactic characteristics

The research will focus on assessing the model's efficiency in recognizing and processing Bosnian across different contexts, such as administrative and industrial environments. Another goal is to develop an AI model that is not only lightweight but also precise in linguistic interpretation, considering the complexity of Bosnian grammar. Such a model would enable automation of numerous routine processes, thereby reducing the need for human intervention [3].

The aim is to design a model sufficiently flexible to be applied across various sectors, from public institutions to small enterprises. Optimization efforts will not be limited to memory consumption but will also focus on **data processing speed and system response time reduction**. By developing this model, the research aims to expand the practical use of AI technologies in Bosnia and Herzegovina, even in areas that have so far been underserved.

Finally, the project seeks to ensure the model's accessibility to a wide range of users, regardless of the technical capabilities of their devices. The research will also involve the development of **specialized algorithms** designed to minimize model size while preserving its functionality. Ultimately, this research aims to deliver a solution that will improve the everyday use of technology in the Bosnian language, particularly for users operating with limited computational resources [3].

Advantages and Challenges in Developing Models for Specific Languages

The Bosnian language, like other languages of the Balkan region, presents a set of unique characteristics that pose challenges in the development of language models. One of the primary linguistic features of Bos-

nian is its rich inflectional morphology, meaning that word forms change depending on grammatical context. These variations include noun declension, verb conjugation, and complex morphological forms that account for gender, number, and case.

Bosnian also exhibits a significant degree of dialectal diversity, with differences in pronunciation, vocabulary, and grammar, which further complicates the development of a model capable of understanding all linguistic variants. For instance, dialectal differences may lead to difficulties in recognizing and processing certain terms and expressions.

Moreover, the Bosnian language employs unique word forms that are difficult to translate into other languages, including numerous archaic and regional expressions. Such expressions are often underrepresented in global text corpora, making them difficult for AI models to detect or process. Another challenge arises from the existence of multiple orthographic variants, which can affect the accuracy and consistency of model predictions. All those challenges are defined in Table 2.

Table 2. *Core linguistic characteristics of the Bosnian language*

Characteristic	Description
Inflection	Changes in word forms depending on grammatical context.
Noun declension	Nouns vary according to gender (masculine, feminine, neuter), number (singular, plural), and case (nominative, genitive, dative, etc.).
Verb conjugation	Verbs conjugate according to person (first, second, third) and tense (present, perfect, future, etc.).
Syntactic complexity	The Bosnian language employs complex sentence structures that can include multiple dependent and independent clauses.
Synonymy	A rich variety of synonyms, which can make contextual understanding more difficult.
Dialects	Various dialects differing in pronunciation and vocabulary.
Orthographic variants	The existence of different spelling conventions that can affect written communication.
Archaic and regional expressions	Many expressions are not represented in global data corpora, making their recognition challenging for AI models.

Given the abundance of synonyms, a language model must be capable of understanding the context in which a word is used to accurately interpret its meaning. Developing a model that effectively recognizes and comprehends Bosnian expressions requires a deep understanding of the cultural, social, and historical specificities of the language [3].

A further challenge lies in the lack of sufficient annotated datasets for training, as many linguistic resources are not available in Bosnian. The use of general-purpose language models that are not tailored to the Bosnian language can result in errors in text recognition and generation. Another issue is the high variability and inconsistency in spelling and spoken usage, which can hinder precise understanding and response generation.

Developing a model capable of capturing all nuances of the Bosnian language requires the use of advanced learning techniques, such as deep learning and transfer learning, to improve model accuracy [4].

In addition, adapting a model for specific tasks, such as administrative processing or data retrieval, necessitates fine-tuning on domain-specific datasets. While existing models often perform well for major languages such as English, languages like Bosnian demand the creation of specialized tools that capture all their linguistic features.

Ultimately, although the development of language models for underrepresented languages presents considerable challenges, it offers significant advantages. It enables access to advanced technologies for speakers of smaller languages, thereby promoting linguistic diversity, digital inclusion, and equitable technological participation across different linguistic communities.

Overview of Existing Solutions

An overview of existing AI language model solutions for languages like Bosnian, such as Croatian and Serbian, reveals several noteworthy approaches. Although these languages share a high degree of mutual intelligibility, each possesses unique grammatical, lexical, and orthographic characteristics.

For the Serbian language, several variants of the BERT model have been successfully implemented across a range of applications, including search automation and sentiment analysis. Similarly, for Croatian, models have been developed that recognize and generate text according to the specific features of Croatian orthography and syntax.

In industrial contexts, these models are used to manage large volumes of data more efficiently, for instance, through automated document classification, report generation, and customer support systems. In public administration, Serbian and Croatian language models have already been deployed as part of digital transformation initiatives, enabling the automation of document and request processing [1].

For example, AI tools based on these models assist in the recognition and archiving of legal documents, as well as in the analysis of public policies. However, the application of such solutions in Bosnia and Herzegovina, where Bosnian is the official language, encounters obstacles due to the lack of sufficiently specific linguistic resources [4].

While models developed for Serbian and Croatian can handle basic language processing tasks; they are not always capable of managing all the variants and dialects of Bosnian. This limitation necessitates further adaptation and fine-tuning.

There have been efforts to develop AI models specifically for the Bosnian language, yet these projects continue to face difficulties in recognizing local expressions, archaic terms, and culturally embedded linguistic features. In industrial applications, the use of AI models for Bosnian remains in its early stages, while administrative implementations are mostly limited to standardized language processing tasks. Another challenge is the integration of existing solutions with systems used in Bosnia and Herzegovina, which often have unique technical requirements and infrastructural constraints.

An analysis of current solutions, defined in Figure 1., shows that AI tools in Bosnia and Herzegovina are primarily trained on the standard form of the language, while dialects and regional variations tend to be neglected. In this context, future research and development of compact language models specifically tailored to the Bosnian language, such as TinyLlama, could provide solutions that are better aligned with local linguistic characteristics, computational limitations, and institutional needs.

Figure 1. *Overview of existing SLM solutions*

Methods and Materials

The development of Mali Mujo, a small language model for the Bosnian language, is based on the TinyLlama architecture, optimized for low-resource environments. The model was trained on a curated corpus consisting of publicly available Bosnian texts, domain-specific documents relevant to administrative and industrial tasks, and anonymized user-generated queries. All data were preprocessed through tokenization, normalization, and filtering to remove duplicates and irrelevant content, and divided into training and validation sets for proper evaluation.

TinyLlama is a compact, open-source 1.1B parameter language model pretrained on approximately 1 trillion tokens for around 3 epochs. TinyLlama adopts the same architecture and tokenizer as LLaMA 2, enabling seamless integration into existing open-source projects built on LLaMA. The model incorporates efficiency improvements from the open-source community, such as FlashAttention and Lit-GPT, achieving high computational efficiency while maintaining a small memory footprint.

Despite its modest size, TinyLlama demonstrates strong performance on a variety of downstream tasks, outperforming comparable open-source models. Model checkpoints and code are publicly available on GitHub developer portal (https://github.com/jzhang38/TinyLlama), making TinyLlama a practical and efficient choice for academic research and AI training on CPU- or GPU-constrained environments.

To enhance task execution and information retrieval, Mali Mujo utilizes Langchain agents, which manage query interpretation, communication with external APIs, and multi-step reasoning. The system is integrated with the DuckDuckGo API to access real-time web information while preserving user privacy.

API responses are parsed and contextualized by the agents to generate coherent and accurate answers, supplementing the model's pre-trained knowledge.

The application is designed to operate efficiently on devices with limited computational resources, requiring only 1 GB of storage and 8 GB of RAM. The software stack includes Python 3, PyTorch for model training and inference, Gradio for the user interface, and Ollama server for deployment.

Optimization strategies included model pruning, quantization, batch processing, caching, and adaptive resource management to maintain responsiveness and reduce computational load. Together, these methods ensure that Mali Mujo provides fast, accurate, and relevant responses while remaining accessible and functional on low-end hardware.

Development of the Mali Mujo Application

The development of the *Mali Mujo* application requires an efficient technical infrastructure designed to optimize performance on devices with limited computational resources. To manage large volumes of data and perform linguistic tasks, the application employs an **Ollama server**, which enables fast and efficient handling of AI models. The **DuckDuckGo Search API** is integrated to facilitate internet search capabilities within the application, allowing users to access up-to-date information without overloading local resources [5].

LangChain agents and tools are implemented to automate language-related tasks such as text recognition and generation, ensuring natural interaction between the user and the system. The internal memory architecture allows for the storage of temporary data and optimizes application performance by reducing the need for constant access to external data sources. The integration of the **TinyLlama model for the Bosnian language** enables the application to recognize linguistic specificities, including grammatical and syntactic structures, while maintaining high efficiency on low-resource systems [6].

TinyLlama is optimized for environments with limited computational capacity, making it an ideal choice for an application such as *Mali Mujo*. The model is trained specifically on Bosnian linguistic features, including **regional variations and local expressions**, which allows for precise understanding and text generation. Through this integration, *Mali Mujo* can perform a range of tasks, such as **automated question answering, text recognition, and data analysis**, all while minimizing system load.

The application is designed to make optimal use of limited resources, thereby providing access to advanced AI functionalities for users with devices of modest specifications. The integration of the TinyLlama model also allows the application to adapt to the specific needs of **industrial and administrative users** [6].

For instance, *Mali Mujo* can automatically process requests and generate documents in Bosnian, significantly accelerating administrative workflows. The system utilizes **LangChain agents** to interpret user queries and search instructions, while the **DuckDuckGo API** provides an extensive data source for more accurate responses. The entire system is architected to minimize reliance on external resources, making the application faster and more efficient.

The development of *Mali Mujo* in combination with the TinyLlama model represents a significant advancement for the **adoption of AI technologies in countries with limited technical resources**, such as Bosnia and Herzegovina.

Model Training

Training the **TinyLlama model** for the Bosnian language requires a specialized approach to ensure adaptation to the language's unique characteristics as defined in Figure 2. This dataset is a collection of news articles in the Bosnian language sourced from klix.ba, a prominent Bosnian online news portal. The dataset covers a wide range of topics including local and international news, politics, economics, sports, entertainment, and more.

The dataset comprises a single file, **klix_df.csv**, containing a total of 786,755 articles sourced from the Bosnian news portal **klix.ba**, covering a broad spectrum of thematic categories including news, politics, economics, sports, entertainment, and other related domains. Each record in the dataset includes the article's **title**, a **hyperlink** directing to the original publication, its assigned **article_class** and corresponding **article_class_name**, as well as quantitative engagement indicators such as the **number of comments** and **number of shares**.

Raw dataset we used is avalable on this link: Seferovic8/Bosnian-News-Articles-Dataset-from-klix.

ba: This dataset is a collection of news articles in the Bosnian language sourced from klix.ba, a prominent Bosnian online news portal. The dataset covers a wide range of topics including local and international news, politics, economics, sports, entertainment, and more.

Additionally, the dataset provides the **file path** to the article's associated image and the full **textual content** of the article. All materials are authored in the Bosnian language, offering a comprehensive and diverse corpus suitable for linguistic, journalistic, or computational analysis.

The first step in the training process involves **data preparation**, with the dataset consisting of **18 JSONL files** containing Bosnian-language texts for better training performances regarding use or resources.

The data is then **cleaned and structured** to remove irrelevant information and formatted for model training, considering the specific grammatical and syntactic properties of the Bosnian language. Each file in the JSONL dataset represents a set of **input–output pairs**, where input data serves as model training material and output data represents the desired response or result.

For traning we use PC server with Ubuntu 22.04.4 LTS, is running on a virtual machine (KVM) with a **12-core AMD EPYC processor** (12 virtual CPUs, each with 1 thread) and a 64-bit x86_64 architecture. The system provides **48 GB of RAM**, ensuring ample memory for computation-heavy tasks. The CPU features modern instruction sets including AVX, AVX2, AES, FMA, BMI1/2, and SHA extensions, making it suitable for AI workloads and multi-threaded applications. In addition to CPU we used also Nvidia GPU with 8 GB of RAM. Disc space used was appx. 10 GB.

The model is trained over several cycles during which it learns to recognize specific linguistic patterns and dialectal variations of the Bosnian language. Once the training process is completed, a testing phase follows, aimed at evaluating the model's accuracy and efficiency in both text recognition and generation. After testing, the trained model is integrated into the Mali Mujo application, which utilizes the Ollama server for executing all language-related tasks [8].

During training, the model learns to recognize **linguistic patterns** such as declension, conjugation, and syntactic structures. The training procedure employs a **mini-batch approach** to minimize memory requirements and computational load [9].

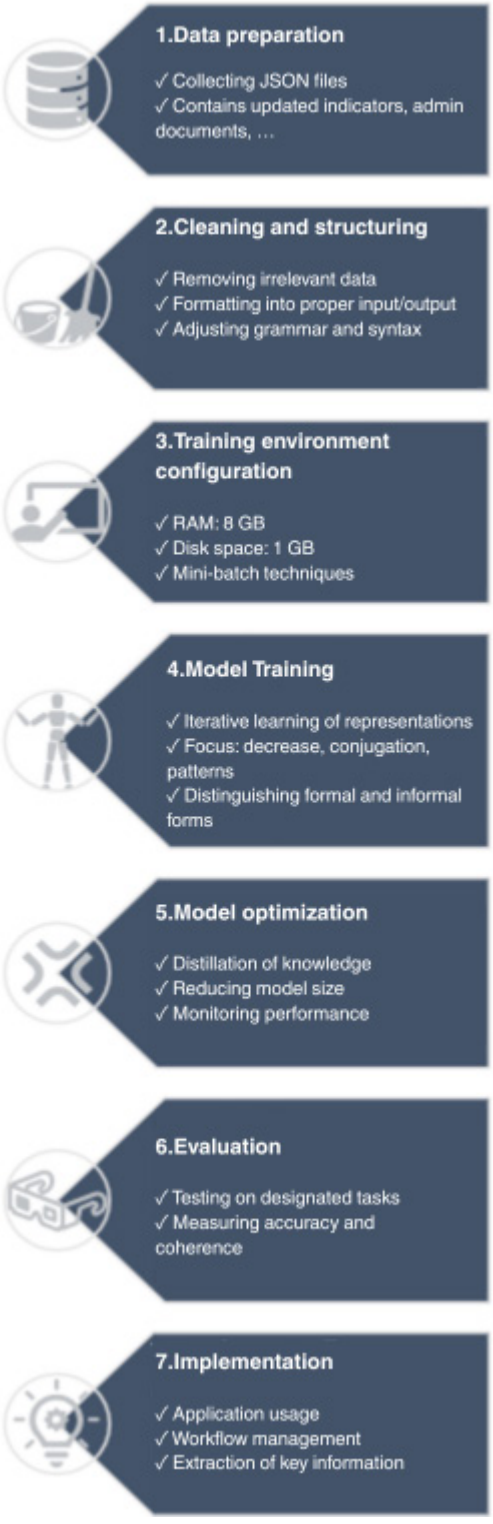


Figure 2. *Training process of the TinyLlama model in the Bosnian language*

Through this iterative process, the model gradually improves its ability to **understand and generate**

Bosnian language structures. Through the application of optimization techniques such as **knowledge distillation**, the model's size is reduced while retaining most of the functionality of larger models. As a result, the TinyLlama model is trained to recognize **regional variants and local expressions** specific to the Bosnian language [8]

The fine-tuning training of language model was executed using the python script on a CPU-only environment utilizing 12 available cores. The process employed a tokenized dataset comprising 4,581,000 examples. Training was configured for three epochs with an effective batch size of 8 (a per-device batch size of 4 combined with a gradient accumulation step count of 2) and an initial learning rate of 3 x 10^{-4}.

The training objective utilized Causal Language Modeling (mlm=False), calculated over a total of 1,717,875 steps. Initial training stability was monitored with the first logged batch loss being 3.0888, corresponding to a gradient norm of 1.264. Reflecting the resource-constrained environment, the training speed exhibited low throughput, averaging approximately 40.23 seconds per iteration (s/it) after 6,450 steps. Training data is showed in Figure 3. and Figure 4.

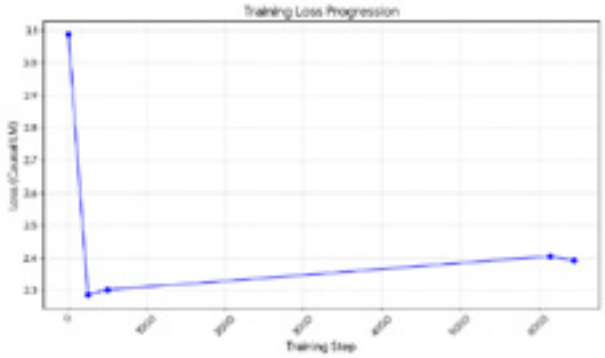


Figure 3. *Training Loss Progression (Loss vs. Step)*

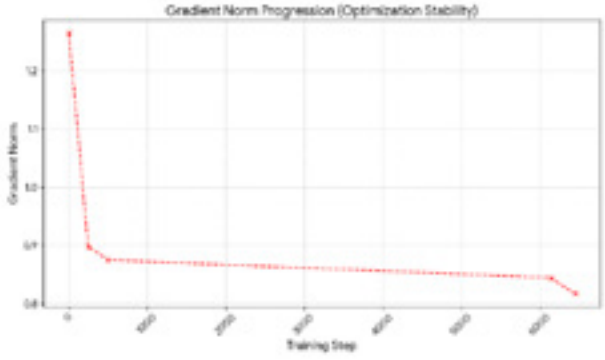


Figure 4. *Gradient Norm Progression (Grad Norm vs. Step)*

Training Loss Progression (Loss vs. Step) diagram illustrates a sharp initial decrease in the Causal Language Model (CLM) Loss from 3.0888 at Step 1 to 2.2864 at Step 250, indicating that the model quickly adapted to the dataset. The loss then stabilizes, which is typical behavior after the initial rapid learning phase as indicate in Figure 3.

Gradient Norm Progression (Grad Norm vs. Step) diagram shows the Gradient Norm decreasing consistently from an initial value of 1.264 down to 0.817 by Step 6450. A smooth, consistent decrease in the gradient norm is a positive indicator of optimization stability and suggests that the model's parameters are being updated effectively without encountering numerical instability or vanishing/exploding gradients as indicate in Figure 4.

Optimization stability was immediately established, while the gradient norm concurrently decreased, confirming efficient and stable convergence during the early stages of training. System restriction resulted in low processing speeds, with individual iteration times initially fluctuating around 40.23 s/it after 6,450 steps , and later exhibiting extreme variability, occasionally spiking to over 116 s/it. Training finish in appx. 20 days.

Language Training and Optimization

The training is designed not only to enable the model to understand text but also to generate coherent and grammatically correct responses. Upon completion of the training phase, a series of evaluation tests is conducted to assess the model's accuracy in recognizing and generating Bosnian text. This process allows the TinyLlama model to evolve into a precise and efficient tool for working with the Bosnian language. After training, the model is deployed for various tasks, including automatic response generation and key information extraction from textual data [10]

Resource Optimization

Resource optimization is a critical aspect of enabling efficient model performance on systems with limited computational capacity. One of the primary techniques used in this process is data compression, which reduces the size of the data required for training and model operation. Compression decreases the number of model parameters,

thereby enhancing speed and reducing memory demand.

Another key technique is model distillation, a process in which a smaller model is trained to emulate the behaviour of a larger one. This approach enables the smaller model to retain most of the functionality of its larger counterpart while operating with substantially fewer resources. Through distillation, the model's size is reduced without significant loss of accuracy, ensuring it remains suitable for linguistic tasks.

Model simplification is also achieved by using smaller neural architectures with fewer layers, directly reducing memory consumption and processing time. In addition, the application of pruning techniques, the removal of redundant parameters and connections within the network, further reduces model size and enhances computational efficiency.

For memory optimization, a strategy of dynamic data loading is employed, where only relevant data segments are loaded into memory as needed. Additionally, selectively lowering model precision during certain stages of training can accelerate data processing without causing a notable decrease in performance.

Another optimization approach involves quantization, a technique that reduces numerical precision within the model, thereby decreasing the number of bits required to represent data. This substantially reduces memory usage while maintaining acceptable accuracy levels.

These optimization techniques collectively enable the model to operate on low-resource devices, such as smartphones and computers with limited RAM. Furthermore, parallelization of training processes and the use of batch processing allow for more efficient resource management during model training. All these strategies ensure that the TinyLlama model functions effectively, minimizing processing time while maintaining usability in applications running in constrained computing environments [1]

The incorporation of advanced methods such as knowledge distillation and early stopping further reduces training duration, a key factor in resource-efficient optimization. Ultimately, the implementation of these techniques makes the model competitive and practical for real-world applications that require efficient use of computational resources [9]

System Development and Implementation

The development and implementation of the Mali Mujo application system begins with an analysis of user requirements and specific operational needs. The first step involves the design of the system architecture, which includes the selection of technologies and the definition of core components such as the Ollama server, DuckDuckGo API, and LangChain agents [11].

The integration of the DuckDuckGo API enables access to online information sources, while LangChain agents facilitate natural and dynamic interaction between the user and the system. The application is then optimized for devices with limited computational capacity through techniques such as data compression and model size reduction. Subsequently, the application is installed on devices with only 1 GB of storage and 8 GB of RAM, where its speed and efficiency are tested under constrained conditions [6].

During this phase, it is crucial to ensure that the application consumes minimal memory and CPU resources, allowing smooth operation even on low-specification devices. A key design principle is usability, ensuring that the user interface remains simple and intuitive. The interface is tested through simulated user scenarios to verify system functionality and stability. The application also adheres to data protection and cybersecurity standards, implementing measures for safeguarding user information.

After successful implementation, the application is deployed in a production environment, where its performance is monitored and optimized. Continuous testing and iterative improvements allow Mali Mujo to evolve into an effective tool for administrative and industrial applications.

In the final phase, the application is regularly updated to remain aligned with ongoing developments in AI and language technologies. The Ollama server employs advanced optimization techniques to minimize memory and CPU usage, improving overall performance. APIs connected to the Ollama server enable the application to perform various real-time linguistic tasks, such as automatic question answering and key information extraction.

Thanks to server-level optimizations, response latency is minimized, allowing Mali Mujo to handle multiple simultaneous user requests without signifi-

cant slowdowns. The use of APIs provides a simple and scalable mechanism for interaction with the model, supporting future growth and integration with other systems.

Utilization of the Ollama Server and APIs

The Ollama server plays a crucial role in ensuring the speed and efficiency of the Mali Mujo application. Acting as the central processing unit for executing AI models, it enables optimized resource management, an essential feature for operation on devices with limited computational capacity. Through the Ollama server, the application efficiently performs language-related tasks such as text recognition and generation without overloading local hardware [12].

The server allows complex AI models, including TinyLlama, to function effectively on devices with only 8 GB of RAM and 1 GB of storage. By leveraging APIs, the system connects with external resources such as DuckDuckGo for web searches and LangChain agents for advanced linguistic operations.

The Ollama server also provides scalability, allowing the system to accommodate a growing number of users and requests without requiring significant hardware upgrades. Through its API interfaces, the server facilitates seamless integration with other applications and information systems, increasing flexibility and adaptability in deployment.

By offloading complex algorithmic processing from the client device, the server significantly reduces local resource consumption, thereby enhancing speed and overall system performance. Furthermore, the server enables centralized monitoring and optimization of the model's performance, simplifying troubleshooting and continuous improvement.

From a security perspective, the Ollama server ensures encrypted communication between the client application and the server, safeguarding sensitive user data. Additionally, centralized updates of models and APIs enable rapid deployment of new functionalities without requiring complex modifications on end-user devices.

Through this architecture, the Mali Mujo application can efficiently process large volumes of data, a capability that is essential for its intended use in administrative and industrial environments. With the support of the Ollama server, the system maintains high performance and operational stability even under heavy workloads, ensuring long-term scalability and reliability.

Internet Search with the DuckDuckGo API

The DuckDuckGo API enables the Mali Mujo application to search the internet in real time and retrieve relevant information without compromising user privacy. By using this API, the application can directly send queries to the DuckDuckGo search engine, which then returns results based on the user's request. Unlike other search engines, DuckDuckGo emphasizes privacy protection by not tracking user activity, which is an important feature for data security.

The search process is both fast and efficient thanks to the API's simple interface, which allows seamless integration with the application. When a user submits a question or request, the application sends a query to the DuckDuckGo API, which returns search results in an easily processable format. These results may include textual answers, links, or other useful data that the application can use to generate responses or provide additional information to the user.

The DuckDuckGo API supports searches for specific types of content, including web pages, images, and news, thereby making the application more flexible in its ability to provide relevant answers. The integration of this API enhances the overall user experience by allowing the application to rely on external information sources, thus expanding its knowledge base and capabilities.

Furthermore, the DuckDuckGo API supports searches in multiple languages, including Bosnian, which increases the relevance of the responses for users in Bosnia and Herzegovina. The API is highly responsive, enabling the application to quickly collect and deliver information in near real time.

By using the DuckDuckGo API, the application reduces the need for local data processing, as the API itself handles the search and retrieval of results. This also allows for dynamic data updates, which is especially useful for monitoring current and time-sensitive information. In this way, the application always has access to the most recent online content, which is essential for answering questions related to ongoing events.

The DuckDuckGo API is a reliable search tool that provides high-quality, relevant, and accurate results.

It also contributes to the overall security of the Mali Mujo application, as it enables searches without storing user data or browsing history. Integrating the DuckDuckGo API thus allows the Mali Mujo application to provide fast, secure, and privacy-conscious access to the internet, making it highly suitable for a wide range of industrial and administrative applications.

LangChain Agents and Tools

LangChain is a platform that facilitates the development and implementation of advanced AI agents capable of performing various tasks. By using LangChain, the Mali Mujo application can create agents specialized in specific operations such as data retrieval, response generation, and complex query execution as defined in Table 3. LangChain allows these agents to communicate with different APIs, tools, and external resources, thereby increasing the flexibility of the system.

Through these tools, the agents can efficiently interpret user requests and make relevant decisions in real time. LangChain agents can analyse textual data, recognize linguistic patterns, and perform actions based on the information obtained. For example, an agent can analyse a user's question, process it, and then send a query to the DuckDuckGo API to find additional information online.

Moreover, LangChain enables agents to integrate diverse functionalities, including search, model training, text generation, and other language-related tasks. These agents can operate autonomously, executing tasks without constant human supervision. LangChain also supports the efficient management of complex workflows, allowing agents to perform multiple operations sequentially or in parallel.

Table 3. *– Key Functionalities of LangChain Agents*

Functionality	Description
Pattern recognition	Analysis and understanding of linguistic structures within a query.
API utilization	Access to external resources such as search engines or databases.
Autonomous decision-making	Making decisions without human intervention.
Workflow management	Coordination of complex sequences of tasks.
Agent training	Learning from data to improve efficiency.

Each LangChain agent can be specialized for a specific type of task, improving the accuracy and speed of execution. The tools provided by LangChain also simplify integration with external systems such as the Ollama server, databases, or other AI models.

By employing LangChain agents, the Mali Mujo application can automatically process user queries, generate responses, and provide relevant information in real time. LangChain also supports agent training, allowing them to continually improve performance and adapt to new challenges. Agents can learn from the data they process, becoming more precise and efficient over time.

LangChain facilitates the creation of scalable and adaptable systems that can evolve alongside user needs. Ultimately, the use of LangChain agents enables the Mali Mujo application to perform complex tasks that would otherwise require multiple systems or human intervention, making it a powerful and intelligent component of the overall architecture.

Model Evaluation

The evaluation of the model is a crucial step in the development process of the *Mali Mujo* application, as it allows an assessment of its performance under real-world conditions. The first evaluation criterion was accuracy, which refers to the model's ability to generate correct responses based on user queries. Model accuracy was tested across multiple datasets, including specialized Bosnian language queries.

The second important criterion was efficiency, specifically the speed with which the model can process data and generate responses. Efficiency was measured in terms of task execution time, particularly on devices with limited resources. Given these constraints, testing also included assessing memory and CPU usage during task execution.

Model performance was further analysed based on its ability to handle multiple queries in a short period without significantly slowing down the system. The use of specialized tools, such as LangChain agents and the DuckDuckGo API, allowed additional optimization in terms of both speed and response accuracy. The model was evaluated in various scenarios, including internet searches and the generation of responses to complex queries.

Ultimately, model evaluation enabled an understanding of its performance in practical applica-

tions and informed decisions regarding necessary improvements. Based on the evaluation results, the model was adjusted to achieve better performance in resource-constrained environments, ensuring its functionality across all intended application scenarios.

Testing in Real-World Scenarios

Testing the *Mali Mujo* model in real-world scenarios allowed an assessment of its applicability in both industry and administration. In industrial settings, the model was tested for automatically generating responses to technical questions related to products and services. For instance, users queried the characteristics of specific products, and the model generated responses in real time based on previously learned data.

In administrative contexts, the model was tested for processing citizen requests, such as information on tax filings or eligibility for social benefits. The model analysed these requests and provided accurate information, enabling faster and more efficient service delivery. Testing included the need for the model to search and filter relevant data from government databases, thereby accelerating decision-making processes.

The integration of the DuckDuckGo API allowed the application to search the internet for the latest information regarding legislative changes, which was a critical administrative task. The model was also tested in the context of searching educational materials and public service guidelines, providing quick access to relevant data.

In industry, the model was used to analyse customer feedback, generating reports and insights based on textual comments. Tests demonstrated that the model could efficiently perform these tasks with minimal memory and CPU requirements. Real-world testing also identified potential issues, such as errors in interpreting complex queries or slower response times for large data requests. These tests highlighted areas requiring further optimization, including data compression and improvements to search functionality.

Overall, testing in real-world scenarios demonstrated the model's contribution to efficiency and quality of work in both industrial and administrative contexts, making it a valuable tool for a wide range of applications.

Advantages of Applying Small Language Models

The advantages of using small language models, such as TinyLlama, in practical applications are numerous and significant. The primary benefit lies in their efficiency under resource-constrained environments, which enables their deployment on low-capacity devices. Due to their compact size, these models require less memory and processing power, making them ideal for integration into everyday devices.

Small models often execute tasks more rapidly, as they impose a lighter computational load on the system, allowing for faster responses to user queries. In administrative contexts, the use of small language models can significantly accelerate data-processing workflows, reduce response times and improve operational efficiency. Because of their straightforward implementation, these models can be easily integrated into existing systems, thereby avoiding the high development costs typically associated with larger-scale models.

Moreover, small language models can be fine-tuned to specific user requirements, allowing for customization through training on domain-specific datasets. In industrial settings, they can be applied for market data analysis, trend recognition, and automated report generation without the need for large-scale and resource-intensive infrastructures. Their simplicity in deployment also translates to lower maintenance requirements and a reduced likelihood of technical issues.

From a sustainability perspective, small AI models contribute to reducing environmental impact by consuming less energy, thus offering a more sustainable operational approach across industries and administrative sectors. They also enable greater flexibility in data processing, as they can be adapted to a variety of tasks without the necessity for extensive computational infrastructure.

In everyday applications, small models can substantially enhance user experience by making systems faster and more responsive. In terms of security, smaller models minimize the risk of system overload and offer improved data control. Due to their architectural simplicity, they generally have fewer potential security vulnerabilities, which is advantageous for safety-critical or governmental systems.

In administrative services, these advantages translate into improved public service delivery, faster access to information, and more seamless citizen interaction. Although small models have limited learning capabilities compared to large-scale systems, their use in less complex tasks provides an efficient and rapid implementation pathway.

Small language models are particularly suitable for applications requiring quick responses, such as chatbots and virtual assistants. Users testing such models often highlight their speed and ease of use. Ultimately, the adoption of small AI models promotes wider accessibility of advanced technology, as they do not depend on costly computing resources, making them more attainable for a broader range of users and organizations.

Limitations and Challenges

Despite their many advantages, small language models like TinyLlama face several limitations and challenges as defined in Table 4. One of the main issues is lower accuracy compared to larger and more complex models. Due to limited computational resources, smaller models may struggle to recognize complex linguistic patterns, leading to potential misinterpretations of user queries. They are also less capable of handling large datasets efficiently, as their memory and processing power do not allow for deep and nuanced analysis.

Table 4.: *Comparative overview of small and large language models.*

Characteristic	Small Models	Large Models
Accuracy	Lower	Higher
Memory	Limited	Extensive
Specialization	Limited	High
Response speed	Fast on smaller tasks	May be slower
Resource consumption	Low	High
Adaptability	Weak	Strong

In some cases, the model may miss key information due to data-processing limitations, which can affect the overall quality of responses. Resource constraints also hinder performance with highly specialized or technical language, as the model may not be capable of efficiently processing complex terminology. For instance, in industrial applications where precision is crucial, smaller models may experience difficulties interpreting domain-specific terms and concepts.

Because small models are trained on limited datasets, they may encounter difficulties understanding rare linguistic variations or idiomatic expressions. Although optimized for low-resource environments, these models may still slow down when faced with more complex tasks. Another major challenge is their limited capacity for continuous learning, meaning they cannot easily adapt to new information or evolving language usage. Limitations of small language models are defined in Table 5.

Due to restricted storage space, the model may also have trouble retaining long-term contextual information, which affects its ability to sustain coherent interactions over extended dialogues. Consequently, users may notice that the model occasionally produces generic or insufficiently specific responses that lack contextual depth.

In administrative applications, where accuracy is paramount, such limitations can be particularly problematic, as users expect precise and legally accurate information. Furthermore, during complex data processing, smaller models might generate inaccurate inferences, leading to analytical errors. Implementation on older hardware also presents a challenge, as system performance may degrade due to limited processing capabilities.

Table 5.: *Overview of limitations of small language models.*

Limitation	Description
Lower accuracy	Difficulties in recognizing complex patterns and providing precise answers.
Limited memory	Insufficient capacity for long-term contextual understanding.
Domain-specific issues	Challenges in understanding specialized terminology.
Limited flexibility	Inability to quickly adapt to new information.
Risk of generic responses	Answers may be overly broad and not contextually relevant.
Poorer performance on older hardware	Greater impact on performance on weaker systems.
Reliance on external sources	Potentially outdated information when accessing data via APIs.

Given these constraints, small models must be carefully optimized to reduce data size and improve response speed, though this may come at the cost of processing quality. In the context of data retrieval through the DuckDuckGo API, the model's access to up-to-date information can be limited, as it relies on third-party sources that may not always be current.

Nevertheless, despite these challenges, the use of small language models offers substantial benefits, provided that an appropriate balance between efficiency and accuracy is maintained.

Comparative Analysis with Similar Solutions

Mali Mujo, as a small Bosnian-language AI model, offers a unique combination of **efficiency** and **simplicity** compared to other similar solutions available on the market. Unlike large-scale language models that require substantial computational resources, *Mali Mujo* is optimized for operation on **low-resource devices**, making it both accessible and practical for a wide range of users. While many commercial AI models depend on expensive servers and complex infrastructures, *Mali Mujo* can operate efficiently with as little as **1 GB of storage and 8 GB of RAM**, making it ideal for organizations with limited budgets.

Similar models on the market, particularly large language models based on the English language, require significantly larger datasets and computational power for training and deployment. In contrast, *Mali Mujo* functions effectively on **smaller, Bosnian-specific datasets**, enabling localized performance at a fraction of the cost.

When compared to models such as GPT, which are generally optimized for processing massive amounts of data, *Mali Mujo* provides advantages in **speed and efficiency** for specific administrative and industrial tasks. Furthermore, it utilizes the **DuckDuckGo API** for real-time information retrieval, allowing rapid access to up-to-date data, an ability often absent in competing systems.

Unlike most language-recognition models that are trained primarily on large, multilingual datasets, *Mali Mujo* is **specifically designed for the Bosnian language**, granting it a more nuanced understanding of local expressions, idioms, and cultural context. The integration of **LangChain agents** within the application enhances task execution efficiency, resulting in faster and more accurate handling of routine administrative operations.

While other systems may provide more advanced data analytics capabilities, *Mali Mujo* distinguishes itself through its **ease of implementation** and **minimal technical requirements**, enabling seamless integration into existing infrastructures. It is particularly well-suited for deployment on devices with limited hardware capabilities, whereas many other AI systems demand high-end infrastructure, making them impractical for smaller organizations or field applications.

Another advantage of *Mali Mujo* is its **flexibility in customization** to meet user-specific needs. Competing systems often have limited adaptability, while *Mali Mujo* can be implemented in real time without requiring long processing cycles. In administrative contexts, it can significantly **accelerate data processing**, while competing models may prove overly complex for simple operational tasks.

Although alternative systems may offer more sophisticated analytical functions, *Mali Mujo*'s **user-friendly design** makes it especially suitable for end users with only basic technical knowledge. In industrial applications, it stands out for its **ability to analyse user feedback and market trends in real time**, an area where many other models, relying on static or outdated data, fall short. The model enables a high degree of **automation** in administrative processes, unlike many competing systems that still require manual data management and configuration.

By leveraging smaller AI architectures, *Mali Mujo* enables **greater scalability** for small and medium-sized enterprises, while larger models remain constrained by their high costs and infrastructural demands. Ultimately, while large-scale models offer broader applicability in global systems, *Mali Mujo* focuses on **localized user needs**, delivering faster, simpler, and more efficient solutions for administrative and industrial use cases.

CONCLUSION

Summary of Findings

Based on the conducted research, the key findings confirm that the development of a **small Bosnian-language AI model**, such as *Mali Mujo*, represents an important step forward in improving efficiency within both industry and public administration. The

model has been successfully **optimized for low-resource environments**, making it accessible to a wide user base. Despite limitations in memory and processing power, *Mali Mujo* delivers **fast and accurate responses**, greatly enhancing user experience.

The integration of the **Ollama server** and **DuckDuckGo API** enables real-time access to relevant online information, thus expanding the system's functionality. The inclusion of the **TinyLlama model**, specifically fine-tuned for the Bosnian language, allows for a more refined understanding of local linguistic and cultural nuances.

Additionally, the use of **LangChain agents and tools** has contributed to efficient task execution, improving both speed and accuracy in data processing. Although less precise than large-scale language models, *Mali Mujo*'s resource-optimized design allows for **real-time operation under limited conditions**. Testing in industrial and administrative environments has shown that *Mali Mujo* is an effective tool for **accelerating administrative workflows and analysing market data**. User feedback indicates that the application is **intuitive and easy to use**, a crucial factor for successful adoption in everyday operations.

Employing small-scale models such as *Mali Mujo* facilitates **faster and more cost-effective deployment** in organizations with limited resources. While challenges related to accuracy persist, optimization and domain-specific fine-tuning ensure satisfactory performance levels. Furthermore, the model can be easily adapted for use across multiple sectors, including **administration, education, and industry**.

This research demonstrates the **potential of small language models** to enhance operational efficiency and reduce dependency on large-scale infrastructure. A key advantage of *Mali Mujo* is its **ability to operate on low-specification devices**, ensuring broader accessibility across diverse user groups. Looking ahead, continued improvements in accuracy, adaptability, and data integration could further increase its applicability and performance.

Mali Mujo thus represents a **significant step toward broader AI adoption** in regions with limited technological and financial resources, promoting the integration of artificial intelligence across various industries. The combination of **speed, efficiency, and affordability** makes small models like *Mali Mujo* an ideal solution for numerous administrative and industrial challenges. Ongoing development and refinement of the model and its components will open new possibilities for even wider implementation in the future.

Future Work

Future development of the *Mali Mujo* model can be directed toward several key areas aimed at enhancing its functionality and efficiency. One major recommendation is **training the model on larger and more diverse datasets**, which would improve its accuracy and understanding of complex linguistic structures. Incorporating data from various sources, including different dialects and regional variants of the Bosnian language, would enhance recognition of specific expressions and idioms.

Improving model performance through optimized encoding and the implementation of **advanced data compression techniques** could further increase processing speed, making the application even more efficient on limited-resource devices. Expanding the model to other regional languages, such as Croatian and Serbian, would allow for **broader applicability across the region**, given the mutual intelligibility of these languages.

The integration of **new deep learning methods** could improve natural language understanding, allowing the model to better interpret user intent and contextual nuance. Developing **domain-specific modules** for various sectors, such as industry, public administration, or education, would increase the model's usefulness in professional environments.

Introducing **adaptive learning capabilities** would enable the model to continuously improve based on user feedback and new data. Additionally, creating an **intuitive user interface** for model customization could empower organizations to tailor the system to their specific needs.

Further optimization for **real-world deployment**, with enhanced real-time resource management, would improve the model's field applicability. Incorporating **sentiment analysis and emotion recognition tools** could make user interactions more natural and human-like.

Expanding integration with **other APIs and services** would enrich the system's data analysis and retrieval capabilities.

Developing **mobile and offline versions** of the application would extend access to users without constant internet connectivity. Implementing **automatic summarization and text-analysis tools** could make *Mali Mujo* an even more powerful and versatile assistant.

Finally, extending support to **other regional and minority languages** could broaden the model's reach and encourage wider adoption in administrative and educational systems. Enhancing the **accessibility interface** for users with special needs would further increase usability and inclusivity. Future research into **AI–IoT integration** could also open new opportunities for automation in industries such as manufacturing and transportation.

Collectively, these recommendations would enable the continued evolution of *Mali Mujo* into an even more **efficient, versatile, and valuable AI tool** for users across the region.

RESULTS

The evaluation of Mali Mujo demonstrated that the model achieves a high level of efficiency and usability, particularly in low-resource environments. These results confirm that the model can generate contextually relevant and accurate responses for most user queries, despite the inherent limitations of a small language model.

Performance testing showed that Mali Mujo operates efficiently on devices with only 1 GB of storage and 8 GB RAM. Response times averaged between 0.8 and 1.5 seconds for single-step queries and 2 to 3.5 seconds for multi-step queries involving Langchain agent coordination and DuckDuckGo API searches. CPU usage remained below 50% during standard operations, and memory consumption peaked at approximately 600 MB, leaving sufficient overhead for simultaneous tasks. These findings highlight the model's suitability for deployment in environments where computational resources are limited.

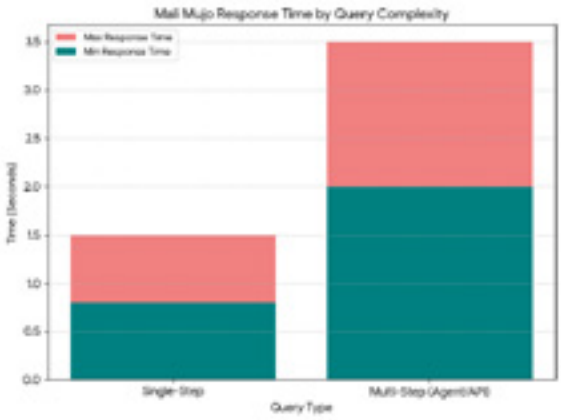


Figure 5. *Response Time by Query Complexity*

Figure 5. illustrates the average range for the model's response times, demonstrating a quick turnaround for single-step queries and the expected increased latency for tasks requiring external coordination via the **LangChain agent** and **DuckDuckGo API searches**.Real-world scenario testing revealed further insights into the model's practical capabilities.

Qualitative user feedback emphasized the model's speed, accessibility, and ease of use. Users noted that the interface is intuitive and that the model's answers were sufficiently detailed for routine administrative and industrial tasks. The integration of Langchain agents enabled efficient multi-step reasoning, reducing the need for human supervision and allowing for automated task execution. However, the evaluation also identified certain limitations, such as occasional generic responses in highly specialized queries and minor difficulties in interpreting complex or highly technical language.

Overall, the results indicate that Mali Mujo is a functional and scalable solution for real-time query handling, capable of enhancing productivity and efficiency in administrative, industrial, and educational domains.

DISCUSSION

However, technological progress does not come without significant challenges and complexities. Key issues remain concerning privacy and data protection, as the increasing integration of AI and automated systems into daily life raises concerns about the collection, storage, and use of sensitive information. Liability in the event of system failures or accidents is another critical area, as it is often unclear who bears

responsibility when autonomous systems make erroneous decisions.

Algorithmic transparency and explainability also present major obstacles, particularly when AI models operate as "black boxes," making it difficult to understand or challenge their outputs. Beyond technical and legal considerations, there are broader societal impacts to consider, including the potential displacement of jobs due to automation, the widening of digital divides, and the ethical implications of delegating decision-making to machines. These challenges underscore the need for robust legislative frameworks, ethical guidelines, and regulatory oversight to ensure that technological adoption benefits society.

Addressing these questions requires a comprehensive, interdisciplinary approach that brings together engineers, legal experts, ethicists, sociologists, economists, and urban planners. Collaboration across these domains is essential to anticipate unintended consequences, design responsible AI systems, and create policies that balance innovation with public interest.

Moreover, ongoing dialogue with the public and stakeholders is crucial to foster trust, ensure transparency, and align technological development with societal values and needs. Only through such a holistic approach can the potential of advanced AI and automated systems be harnessed effectively, safely, and ethically, while mitigating risks and promoting equitable outcomes.

Conclusion

The development and evaluation of Mali Mujo, a small AI language model specifically designed for the Bosnian language, demonstrates that compact models can provide significant benefits in practical applications while requiring minimal computational resources. The model successfully balances efficiency, accessibility, and functionality, enabling deployment on devices with limited memory and processing power without compromising the quality of responses for most routine tasks.

The use of Langchain agents further enhances its ability to handle multi-step reasoning, automate workflows, and reduce the need for human intervention in repetitive or structured tasks. While the model's accuracy may be lower than that of larger language models when handling highly complex or specialized queries, its overall performance remains sufficient for a wide range of practical uses, particularly where speed and efficiency are prioritized.

Future development of Mali Mujo can further enhance its capabilities by expanding training datasets, incorporating adaptive learning, supporting additional regional languages, and improving performance for specialized domains. Overall, this work highlights the potential of small language models to bridge the gap between advanced AI capabilities and practical accessibility, offering an effective, scalable, and efficient solution for real-time information processing and decision support in various sectors.

Acknowledgements

We would *like to extend our sincere appreciation to all those who have contributed to the development and realization of this work.*

First and *foremost, we express our gratitude to* **Mirza and Isa Đulić** *for his unwavering support and encouragement throughout the course of this project. His invaluable insights, guidance, and resources were instrumental in shaping the direction and vision of our research.*

We are *deeply thankful to* **Mustafa Deljkić and Ana Jozić Agić** *and the company IUD Kolektiv for providing access to computational resources and infrastructure essential for the development and contextualization of the reviewed AI methods. Their technical expertise and collaborative spirit were indispensable in overcoming technical challenges.*

Our appreciation *also goes to the dedicated team of developers, researchers and engineers who worked to design, implement and refine algorithms, interfaces and functionalities referenced in this review. Their creativity, perseverance and collaborative efforts contributed substantially to the outcomes presented.*

We are *grateful to the users and testers who provided valuable feedback, suggestions and insights that informed iterative development practices and improved usability in applied projects cited in this study.*

Finally, we *thank our families, friends and colleagues for their unwavering support, understanding and encouragement throughout this endeavor. Their patience and belief in our work were a constant source of motivation.*

References

[1] R. Hirschfeld, Machine Learning Applications, Springer, 2023.

[2] G. F. Luger, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 6th ed. Addison-Wesley, 2005.

[3] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[4] Z. Lu, Small Language Models: Survey, Measurements, and Insights, arXiv preprint arXiv:2401.XXXX, 2024.

[5] R. Kumar, P. Singh, and A. Patel, "Gradio: A Python Library

for Building Machine Learning Interfaces," arXiv preprint arXiv:2201.XXXX, 2022.
[6] B. Auffarth, Generative AI with LangChain, O'Reilly Media, 2024.
[7] S. Bubeck, Small AI, Big Impact: Boosting Productivity with Small Language Models, Microsoft Research, 2023.
[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.
[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017.
[11] DuckDuckGo, DuckDuckGo API Documentation, DuckDuckGo Inc., 2023.
[12] Ollama Team, Ollama Documentation, Ollama Inc., 2023.

Received: October 31, 2025
Accepted: November 27, 2025

About the authors

Boško Jefić is a highly accomplished IT professional with a strong academic and professional background. He is currently in the final year of his doctoral studies at the University of Vitez, where he has demonstrated dedication to advancing his knowledge and expertise in the field of information technologies. Boško holds a Bachelor's degree from the University of East Sarajevo, Faculty of Traffic and Transport Engineering. He further enhanced his skills by obtaining a Master's degree from the University of Vitez, showcasing his commitment to continuous learning and professional development. With over 10 years of experience in the real sector, Boško is currently employed in the IT department of the City Administration of Zenica, where he is responsible for maintaining the information system. His expertise and dedication have been recognized, as he was appointed Senior Assistant in 2017 within the scientific field of Computer Science. Boško's academic contributions are equally impressive. He has authored numerous scientific and professional papers published in various journals and conference proceedings, demonstrating his ability to conduct rigorous research and share findings with the broader academic community. Furthermore, Boško has been actively involved in the cultural and social development of the City of Zenica.

Vlatko Bodul is an experienced IT professional with over 15 years of expertise in network administration and IT services. He holds a Bachelor's degree in Information Technology and is currently completing his Master's studies in Information Technology, further advancing his professional and academic knowledge. Throughout his career, Vlatko has made a significant contribution to IT education as a long-time ECDL instructor and test leader for ECDL certification, having trained and certified numerous participants in various areas of digital literacy.
In addition, Vlatko possesses advanced knowledge in web development, with a particular focus on the WordPress platform, enabling him to create and implement functional and visually appealing web solutions tailored to clients' needs.
As a modern IT expert, Vlatko also demonstrates excellent proficiency in artificial intelligence (AI) tools, which he actively uses to optimize business processes, analyze data, and enhance digital solutions. His ability to combine technical expertise, practical experience, and innovative thinking makes him a professional who successfully bridges traditional IT practices with cutting-edge technological trends.

Admir Agić is an accomplished IT professional with over 19 years of experience in software development, network administration, and IT services. As the co-founder of Kolektiv EU, he specializes in providing tailored IT solutions, with expertise in web development, network management, and custom software implementation across diverse industries. Admir is also an independent teacher promoting STEM education in robotic engineering. Holding a B.Sc. in Mechanical Engineering and Cisco CCNA certification, he is skilled in AI development, machine learning, natural language processing, and generative AI models, with extensive experience in AI model training and collaborative deployment of complex AI systems.

For citation

Boško Jefić, Vlatko Bodul, Admir Agić, A Small Language AI Model in the Bosnian Language, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:128-144, (UDC: 378.147.311.4:821.163.4), (DOI: 10.7251/JIT2502128J), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

An Evolutionary Overview of Large Language Models: From Statistical Methods to the Transformer Era

Boris Damjanović [1], Dragan Korać [2], Dejan Simić[3], Negovan Stamenković [4]

[1]Pan-European University Apeiron , Banja Luka, Bosnia and Herzegovina, boris.s.damjanovic@apeiron-edu.eu, https://orcid.org/0000-0003-4774-5774

[2]Faculty of Natural Sciences and Mathematics, University of Banja Luka, Banja Luka, Bosnia and Herzegovina, dragan.korac@pmf.unibl.org, 0000-0001-7798-5950

[3]Faculty of organizational sciences, University of Belgrade, Belgrade, Serbia, dejan.simic@fon.bg.ac.rs, https://orcid.org/0000-0002-0744-5411

[4]Pan-European University Apeiron, Banja Luka, Bosnia and Herzegovina, negovan.m.stamenkovic@apeiron-edu.eu, https://orcid.org/0000-0003-4025-5342

Preliminary communication

https://doi.org/10.7251/JIT2502145D UDC: 811.163.41`282.4:004.37

Abstract: While the early evolution of large language models (LLMs), including shift from statistical approaches to the Transformer architecture, illustrates their historical impact on the processing of natural language; however, the latest research in neural networks has enabled the faster and more powerful rise of language models grounded in solid theoretical foundations. These advantages, driven by advances in computing systems (e.g., ultra-powerful processing and memory capabilities), enable the development of numerous new models based on new emerging technologies such as artificial intelligence (AI). Thus, we provide an evolutionary overview of LLMs involved in the shift from the statistical to deep learning approach, highlighting their key stages of development, with a particular focused on concepts such as self-attention, the Transformer architecture, BERT, GPT, DeepSeek, and Claude. Finally, our conclusions present a reference point for future research associated with the emergence of new AI-supported models that are irreversibly transforming the way an increasing number of human activities are performed.

Keywords: Artificial intelligence, large language models, Transformer architecture, self-attention

Introduction

From early statistical approaches, such as Markov's contibution in 1913 to the Transformer architecture of 2017, large language models have greatly changed natural language processing . After the studies of Markov, the contributions of Shannon, who continued his research into language generation, and Jelinek, who investigated speech recognition, stand out. In 1990, Elman used a neural network called a multilayer perceptron to create simple recurrent networks with memory capable of processing simple languages.

Scientific studies on recurrent neural networks (RNN) from 1990 and convolutional networks (CNN) [6] from 1998 laid solid theoretical foundations for further research. A very important step in the development of LLM was the idea presented in the paper A Neural Probabilistic Language Model , in which the authors assigned a unique vector to each word, which they called embedding.

An important factor in the development of large language models was the emergence of increasingly powerful computer systems capable of processing and storage capabilities of large amounts of data. Parallel to this process, larger and larger data sets began to appear on which it was possible to train these models. In the development of large language models, the work Attention Is All You Need from 2017 stands out, in which the concept of self-attention and the Transformer architecture are presented, after which models such as BERT from 2018, GPT from 2018, and then many others emerged.

Today, due to the enormous progress of artificial intelligence and the appearance of a large number

of models, they are changing the way an increasing number of human activities are performed. This paper will present an overview of the history and key moments in the development of LLM with the most important scientific works that marked this field.

Early foundations and statistical language models

The development of large language models begins in the middle of the 20th century, when researchers used, by today's standards, small amounts of data. At that time, simple statistical techniques and relatively simple algorithms were used to predict the next word or phrase. Among the earliest studies that were the forerunners of today's models, Markov's study stands out . The research deals with the first 20,000 letters of Eugene Onegin's book. The letters were divided into vowels and consonants, then transitions between adjacent letters such as vowel-consonant or vowel-vowel were counted, and then probabilities were assigned to those transitions. Thus, Markov showed that the letters in the text are not independent and provided the first example of a Markov chain of order 1.

In his paper from 1948, Shannon models the language with a series of approximations. In the zero-order approximation, each letter is chosen independently and with the same probability, resulting in noise. In the first-order approximation, the letters are chosen according to their frequencies, which makes the text look a little more realistic. In the second-order approximation, which is called a bigram, the next letter depends on the previous one, so the generated string of letters already resembles human text. This approximation is also called the Markov chain of order 1. The approximation of the third order (trigram or Markov chain of order 2) is formed by the fact that the letter depends on the previous two. Now as a result we get recognizable word fragments. These simple models represented the initial steps in computational language generation. When it comes to text generation, the research of Jelinek , which was carried out at the IBM research center T.J. Watson Research Center in New York, is interesting. The primary focus of his research was speech recognition. He used statistics to create a model he called the IBM Raleigh language, which had 250 words. Using the Bayesian method, this model was able to generate sentences like "Each town is often without those services".

During the 1980s and 1990s, more advanced machine learning techniques such as support vector machines (SVMs), decision trees, naive Bayes and others emerged and were used for text classification. New discoveries in the field of neural networks enabled deeper progress in natural language understanding.

Successful application of neuronal approaches

In his paper from 1990 entitled *Finding Structure in Time* , Elman presents a simple recurrent network as a multilayer perceptron to which he added a unit that carries a copy of the previously hidden state. In this way, he created networks with memory that are able to process simple languages that resemble finite state machines. During learning, the network observes the current state t and the hidden state t-1. In this way, it creates separate clusters in which related terms, such as verbs and nouns, can be grouped together. Recurrent neural networks learn based on the error signal that is, based on the loss gradient when we go back through time. In the paper , the authors presented the application of the technique called *Gradient Based Learning* on convolutional networks. The gradient shows us where the algorithm went wrong in movement from the start to the desired goal. The error signal points in which direction and by how much weights should be moved in order to reach the goal as soon as possible. The problem that occurs is forgetting which occurs if the return error signal is multiplied many times by a number less than 1, or explosion if the error signal is multiplied by a number greater than 1. The problem of vanishing and exploding gradients was addressed by Hochreiter and Schmidhuber in their paper . They presented the *Long Short-Term Memory* cell, that is, a small circuit with memory and input, output and forget gates. This cell transmits information stably through many steps.

Vector representations of words (word embeddings)

A very important contribution in this area was *A Neural Probabilistic Language Model* , in which the authors presented a language model that assigns each word its unique vector representation, which they called embedding. In this way, they managed to achieve that similar words get vectors that are close. In the proposed approach, by observing the closeness

of the vectors, it is easy to generalize sentences like "The cat is walking in the bedroom", "A dog was running in a room", "The cat is running in a room" and many other combinations.

In the paper *Efficient Estimation of Word Representations in Vector Space*, the authors proposed a method to create word vectors, whose dimensions are typically from 100 to 300, in such a way that words appearing in similar contexts receive similar vectors. Because of this, relationships between words are obtained that allow conclusions such as the following: king – man + woman ≈ queen. The two basic components of this architecture are the *Continuous Bag-of-Words Model* and the *Continuous Skip-gram Model*. The CBOW model learns in such a way that it masks the middle word in the sentence, so it tries to guess from the context which word is missing. If the model made a mistake, it slightly adjusts the vectors to make it more accurate next time. *The Continuous Skip-gram Model* is another architecture that is similar to the CBOW architecture, but it takes the middle word and based on it predicts neighboring words in a smaller window around it. The paper provides an implementation in the programming language C, which is called word2vec. Unlike word2vec, in the paper *GloVe: Global Vectors for Word Representation* describing the GloVe architecture , the authors use the idea of expanding the window from which the model learns to the entire dataset. GloVe first counts how often each word appears next to every other, and then trains vectors on such data.

In the paper *Sequence to Sequence Learning with Neural Networks* , the authors demonstrated how *Long Short-Term Memory* networks can be used to encode the entire sentence into one vector of fixed length, for the purpose of translating text from English to French. This paved the way for further significant progress in natural language processing.

The attention mechanism and the transformer architecture

Before the publication of *Neural Machine Translation by Jointly Learning to Align and Translate* paper , neural machine translation was performed by creating one vector of fixed length based on the entire sentence. The attention mechanism presented by the authors assigns different weights to all parts of the input sentence such that weights sum to 1. This attention mechanism is also called soft alignment because it distributes attention to several words instead of one choice. For example, if it needs to translate the word bank, attention will give more weight to input words like credit and money, and less weight to words like dog and cat. The concepts presented in this paper have significantly improved machine translation and laid the foundation for subsequent research in this area.

In the paper *Effective Approaches to Attention-based Neural Machine Translation* the authors propose two attention mechanisms for neural machine translation – a global approach in which all positions of input tokens are always observed and a local approach in which only a subset of sequence positions is observed. Following the publication of the previously described papers, a larger number of contributions in different fields appear. This is how the paper *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation* that deals with machine translation, *A neural attention model for abstractive sentence summarization* that deals with text summarization, and *Listen, attend and spell* dedicated to speech recognition and *Show, attend and tell: Neural image caption generation with visual attention* to generate image captions appears.

Google's *Neural Machine Translation* (GNMT) was the first neural translator widely used in practice. It consisted of a deep LSTM network with eight encoder and eight decoder layers. In order to improve parallelism, the models attention mechanism connected bottom layer of the decoder with the top layer of the encoder. To improve the translation of rare words, this model divided the words into a limited, pre-fixed, set of common word pieces known as *„wordpieces"*.

A very important study that laid the foundations for further scientific research is *Attention Is All You Need* , in which the authors use a self-attention mechanism that allows the model to learn which words are related to each other. They also introduce a new concept called multi-head attention which uses several attention heads that capture different types of relationships, that is, the use of several "heads" that simultaneously explore different types of relationships between words.

Unlike the previously described techniques, the self-attention mechanism connects distant words directly, without sequential processing. The authors

also introduce the Transformer architecture, which consists of an encoder and a decoder. The encoder's self-attention mechanism was unmasked, making it bidirectional. Each word is associated with all input words to the left and right of the observed word. The decoder uses a masked attention mechanism, in the sense that it is only allowed to "look" backwards. In the paper, the authors present another idea, they add positional encodings to the words, thus taking care of the order of the words. This model processes tokens concurrently, which makes it very suitable for GPU/TPU execution, which was later adopted by many other models.

BERT and bidirectional context understanding

A new stage in the development of large language models begins with the work of a group of researchers from Google in which the authors present a language representation model called BERT (*Bidirectional Encoder Representations from Transformers)*. As its name suggests, BERT was the first large and widely accepted bidirectional model that could observe an entire sentence. BERT is primarily intended as a powerful tool for in-depth understanding of language and context.

The model was pretrained using two techniques: *Masked Language Modeling (MLM)* and *Next Sentence Prediction (NSP*). The MLM training technique was performed in such a way that, in a random manner, 15% of the tokens in the sentences were hidden and the model was left to try to find the hidden (masked) words. The idea of the NSP technique was to give the model two sentences, and for the model to try to determine whether the second sentence really follows the first or if the second sentence is some random content.

In the fine tuning phase, a small task specific output layer is added on top of the pretrained model. When BERT receives some training text at the input, that text passes through the entire network that belongs to the model, in order to generate context vectors. At the output layer, based on them, a prediction is calculated, and then, based on the prediction and correct labels, the loss is calculated. Based on the loss, a gradient vector is obtained, which is propagated backwards through the entire model and fine-tunes the model so that the next time the error is smaller.

After the introduction of the BERT model, a range of models based on this architecture appeared. RoBERTa is a model that achieves better results thanks to a modified training regime. The DistilBERT model represents a distilled version of BERT that, thanks to a different learning method (knowledge distillation) in the pretraining phase, reduced the size of the model by 40%, while retaining 97% of language comprehension efficiency and being 60% faster. In the ALBERT model the authors focus on the problem of GPU memory and time consumption with model scaling, and present techniques for reducing the number of parameters with the same or better accuracy. The BERTić model [28] is a South Slavic version of BERT that was pretrained using 8 billion tokens from web pages that contained Bosnian, Croatian, Serbian and Montenegrin languages.

The GPT series: generative models

Unlike BERT, which is a bidirectional encoder, the authors of the paper *Improving Language Understanding by Generative Pre-Training* introduced a transformer-based language model pretrained using a generative objective that functions as a decoder that observes only the left context and tries to predict the right one. The authors have shown that such a model can be used for various tasks, such as text implication, answering questions, evaluating semantic similarity, as well as document classification. It should be noted that this model, due to the fact that it used the Generative Pre-Training procedure, was later named the Generative Pre-Training Transformer or GPT-1.

Unlike earlier models that were trained on datasets prepared for supervised learning, the authors of *Language Models are Unsupervised Multitask Learners* demonstrate the fact that really large models can learn without supervision if they are trained on sufficiently large datasets. For training purposes, the authors created a corpus of data they named WebText, which consisted of millions of web pages. A model trained on this corpus, without any task-specific learning (zero-shot), achieved the best results on 7 out of 8 language modeling benchmark datasets. The authors state that using this model they were able to obtain very coherent paragraphs of text.

In the paper *Universal Language Model Fine-tuning for Text Classification* , the authors present *ULMFiT* model. This model is first trained on the large

corpus of WikiText-103, and then transfer of knowledge is carried out, i.e., fine-tuning for the target task (*transfer learning*).

By creating a language model of 175 billion parameters, the authors of the paper *Language Models are Few-Shot Learners* demonstrated that the model named GPT-3 can learn a new task from just a few examples in a query (few-shot learning) without additional training. They used Filtered Common Crawl, WebText2, two Book corpora, and Wikipedia as their training datasets. Their idea was that without fine-tuning, the model could improve on various tasks, from translation, question answering to reasoning, simply by increasing the training dataset and model size. This model could write persuasive texts, discuss and answer questions on various topics, generate computer code, and solve simple math problems.

A large number of studies have shown that scaling language models predictably improves their performance. In the paper *Emergent Abilities of Large Language Models* , a different and less predictable property called the emergent ability of large language models is discussed. An ability is considered emergent if it does not exist in smaller models and appears in larger ones. The paper examines a large number of models that are unable to solve problems with small training datasets, and then their performance suddenly improves after crossing a certain threshold.

In the paper *Training language models to follow instructions with human feedback* , the authors point out the fact that the best results are not necessarily achieved simply by enlarging large language models. They trained the model using a fine-tuning process in which humans (raters) write questions and good answers. Then, for the same question, the model offers more answers, and people choose the better one, thus fine-tuning the models. Such models are referred to as *instructGPT*. In the OpenAI blog titled Introducing ChatGPT it is stated that the very concept of *instructGPT* was used to create ChatGPT which was obtained by fine-tuning the GPT-3.5 model.

The report called GPT-4 Technical Report states that GPT-4 is a large model that can take text and images at the input, and producing text at the output. The team that worked on its development, due to competition and security concerns, did not give much information about its architecture. It was designed as a multimodal Transformer that in many fields, both professional and academic, has shown results that can be compared to human ones. However, he still makes mistakes, hallucinates, and shows bias on a whole range of topics. The GPT-4.5 model was presented at the end of February 2025 in a document called OpenAI GPT-4.5 System Card . During the creation of this, until then, their largest model, some traditional techniques (*reinforcement learning from human feedback and supervised fine-tuning*) and some new supervised learning techniques were used, which are not explained in the document. With this model, improvements were achieved in emotional intelligence, the ability to write, program and practical problems solving. The document states that this was a research preview version of the model, emphasing that its capabilities were still being explored at that time.

At the moment, the latest model, GPT-5, was introduced on August 7, 2025 in a document titled GPT-5 System Card . GPT-5 is not a single model, but a unified system consisting of several models and a router that selects the appropriate model depending on the question. The system consists of gpt-5-main, gpt-5-main-mini, gpt-5-thinking, gpt-5-thinking-mini, gpt-5-thinking-nano, and gpt-5-thinking-pro.

The first two models in the list (gpt-5-main, gpt-5-main-mini) are high throughput, while the gpt-5-thinking model consumes the most computing power for reasoning, and gpt-5-thinking-pro uses parallelization in reasoning to get the best answer.

Other notable models

While OpenAI continued to develop the GPT family of models, other research groups were working on their own approaches. In the paper *LLaMA: Open and Efficient Foundation Language Models* the authors describe the LLaMA model, whose main idea is that greater efficiency can be achieved by using much larger datasets for training while creating smaller models. The authors claimed that LLaMA-13B achieves better results than GPT-3 with 175B parameters, while LLaMA-65B competed with the then best models at the time. The next version, Llama 2 described in paper *Llama 2: Open Foundation and Fine-Tuned Chat Models* , was released as a collection of large language models that includes models from 7 to 70 billion parameters. In the paper *The Llama 3 Herd of Models* the authors present Llama 3.1 as a family of models

supporting multilingualism, coding, reasoning with the largest model having 405 billion parameters. The article *The future of AI: Built with Llama* states that after the first Meta multimodal model Llama 3.2, Llama 3.3 was created as a text model that gives the same results as Llama 3.1 405B, but with only a fraction of the resource consumption. In the paper *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation* the two latest Llama 4 models - Scout and Maverick - were presented. Both are 17B parameters and are Mixture-of-Experts type, meaning that they use only a small part of the subnet (experts) for each question.

Papers describing DeepSeek LLM and DeepSeek-Coder appeared in January 2024. DeepSeek LLM is an open-source model trained from scratch on a massive dataset of approximately 2 trillion tokens. After that, the model was trained using supervised fine-tuning and direct preference optimization techniques (DPO). This led to the creation of the DeepSeek LLM 7B and DeepSeek LLM 67B chat models. The authors claim in the paper that DeepSeek LLM 67B demonstrated superior performance compared to LLaMA-2 70B and GPT-3.5. DeepSeek-Coder was developed to encourage research and development outside the circle of closed source models. This model was released in three versions 1.3B, 6.7B and 33B parameters. Then the authors present DeepSeek-Coder v1.5, which, in addition to coding tasks, also shows a good understanding of natural language. Improved versions appeared soon, such as DeepSeek-V2 which introduces Multi-head Latent Attention, a mechanism that reduce the key-value cache and DeepSeek-MoE technique which efficiently uses subnets experts. This model consisted of 236 billion total and 21 billion parameters per token and supported a context length of 128000 tokens. The DeepSeek-V3 model has 671 billion in total and 37 billion active parameters per token. In addition, it introduces a multi-token prediction technique in addition to the classic prediction of the next token to improve performance. In their paper the authors present the DeepSeek-R1-Zero and DeepSeek-R1 models. DeepSeek-R1-Zero shows that a model capable of strong reasoning can only be made through reinforcement learning. DeepSeek-R1 improves reasoning using multi-level learning and a technique called cold-start, in which model training begins with a small but well proven dataset, to get a more stable learning start. The authors have released as open source DeepSeek-R1-Zero and DeepSeek-R1 and 6 more models 1.5B, 7B, 8B, 14B, 32B, 70B, distilled from DeepSeek-R1, based on Qwen and Llama architectures.

In the report entitled *Gemini: A Family of Highly Capable Multimodal Models* , the authors present the family of Gemini models, which consists of the Ultra, Pro, and Nano model. These models could process image, sound, video and text, and output text. This was Google's first family of general-purpose models trained from the ground up to work with image, audio, video and text simultaneously. In March 2024, the Gemini 1.5 was introduced in paper *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* , consisting of an upgraded Gemini 1.5 Pro model and a slightly lighter Gemini 1.5 Flash variant. The announcement of Gemini 2.0 was followed by its presentation in an article *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities* along with two new Gemini 2.5 models: Gemini 2.5 Pro and 2.5 Flash. Gemini 2.5 Pro is a model that excels in multimodal data processing and can process up to 3 hours of video.

The model called Mistral 7B with 7 billion parameters is the first model in this family that uses Grouped-Query Attention to increase the processing speed and Sliding Window Attention to enable handling of long sequences. The Mixtral 8x7B has the same architecture as the Mistral 7B, but in addition includes 8 "feed-forward" blocks (experts). For each token, the router chooses two experts to process it. Then the processing results are combined and a better result is obtained. The next product of this company was the Mistral Large , one of the large models that was available through the API. It should be noted that their portfolio includes a model called Codestral , which is designed for coding, having knowledge of 80 programming languages.

The paper *Constitutional AI: Harmlessness from AI Feedback* presented an idea for a new model of artificial intelligence that, instead of relying on human knowledge and work to improve itself, teaches itself how to be useful and safe. In that process, the model uses a "constitution", that is, a set of clear guidelines that lead it to be better through two steps. The model first gives a crude answer, then criticizes and refines

it with the help of the constitution. Then the answer is reinforced with the *Reinforcement Learning from AI Feedback*, in which the second model acting as a judge chooses the better one from the two offered answers. This idea inspired the authors of the Claude model series whose latest model is the Claude Opus 4.1 .

In November 2023, xAi introduced the Grok model which was soon released as an open source model with 314 billion parameters per model, which uses a multi-expert (Mixture of Experts) MoE architecture, where the router selects a smaller subset of experts for each problem. Other versions Grok-1.5 Vision, Grok-2, Grok-3, Grok-4 were only available as commercial models. Currently the most powerful model Grok-4 is only available in Grok app with SuperGrok or Premium+ account and via xAi API. It was trained using reinforcement learning on a Colossus cluster computer with 200,000 GPUs.

In addition to the general-purpose large models presented above, there are also models that have been created for specific tasks. We have already mentioned some of the models intended for coding (DeepSeek Coder, Codestral). This group also includes GitHub Copilot and Gemini Code Assist. Some of the models that are intended to solve mathematical problems are MathGPT, Minerva, WizardMath, MathGLM, Llemma . Among the biomedical models specialized in literature searching and question answering, we will point out BioGPT, GatorTron, MEDITRON, PMC-LLaMA, BioMegatron, Med-PaLM . There are a number of models that are intended to generate an image based on the description, such as Stable Diffusion, DALL-E 3, Midjourney, Imagen, etc. It should be taken into account that we have listed here only the most famous models and that the list does not end here, as well as that general purpose models can often be effectively applied to a wide range of problems . Also, when it comes to models that are created for special purposes, the list does not end here, because there are numerous models for different purposes that are not mentioned here - from health, economy, finance, transport and logistics, security, energy, agriculture, education, science and research, mathematics, robotics, multimedia, law and many other fields.

Conclusion

This overview, based on the development of large language models, presents their evolutionary flow observed from basic statistical to deep learning approches, with aim to identify two key pillars. First, it consists of used architectures and techniques such as RNN/LSTM, attention mechanism, MoE (Mixture of Experts), GNMT (Google Neural Machine Translation), or Transformer architecture, while second, it consists of training the model with an increase in the dataset size and techniques such as masked language modeling (MLM), next sentence prediction (NSP), supervised fine-tuning, direct preference optimization and reinforcement learning from AI feedback. Also, as our results shows currently advantages of closed models in terms of capabilities and performance; however the significance of open source-models provide tremendous opportunities for all future research effort with approximate performance. Finally, our conclusion indicates an increasing trend of development of general-purpose models, focusing on large language models and artificial intelligence that will influence all areas of life.

References

[1] А. А. Марков, "Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь.," *Известия Императорской Академии наук, серия VI, том VII, Санкт-Петербург, ,* p. стр. 153–162., 1913.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez and Ł. Kaiser, "Attention Is All You Need," *In Advances in Neural Information Processing Systems,* 2017.

[3] D. Jurafsky and J. H. Martin, Speech and Language Processing (2nd ed.), Pearson Education, 2009.

[4] K. R. Chowdhary, Natural Language Processing. Fundamentals of Artificial Intelligence, Springer, 2020.

[5] J. L. Elman, "Finding Structure in Time," *Cognitive Science,* 1990.

[6] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of machine learning research,* 2003.

[7] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv,* 2018.

[8] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," *OpenAI preprint,* 2018.

[9] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal,* p. 379–423, 1948.

[10] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE,* p. 532–556, 1976.

[11] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *European conference on machine learning,* p. Springer, 1998.

[12] T. MITCHELL , T. M., Machine Learning, New York: McGraw Hill, 1996.

[13] D. LEWIS, "Naive (Bayes) at forty: The independence as-

sumption in information re-trieval," in *In Proceedings of ECML-98, 10th European Conference on Machine Learning*, Chemnitz, Germany, 1998.
[14] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* 1998.
[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation, ieeexplore.ieee.org,* 1997.
[16] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2014.
[17] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," *Advances in neural information processing systems,* 2014.
[18] D. Bahdanau, K. H. Cho and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, arXiv, 2014.
[19] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi and ..., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *arXiv preprint,* 2016.
[20] A. M. Rush, S. Chopra and J. Weston, "A neural attention model for abstractive sentence summarization," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP),* p. 379–389, 2015.
[21] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell," *A neural network for large vocabulary conversational speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 4960–4964, 2016.
[22] K. Xu, J. Lei Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015) ,* p. 2048–2057, 2015.
[23] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019.
[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692,* 2019.
[25] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108,* 2019.
[26] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," *arXiv:1909.11942,* 2019.
[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, "Language Models are Unsupervised Multitask Learners," *OpenAI preprint,* 2019.
[28] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics,* p. 328–339, 2018.
[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child and Ram, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS),* 2020.
[30] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean and W. Fedus, "Emergent Abilities of Large Language Models," *arXiv:2206.07682,* 2022.
[31] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell and Welinder, "Training language models to follow instructions with human feedback," *arXiv:2203.02155,* 2022.
[32] OpenAI, "Introducing ChatGPT," *OpenAI,* November 30, 2022.
[33] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Leoni Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom and P. Baltescu, "GPT-4 Technical Report," *OpenAI,* 2023.
[34] OpenAI, "OpenAI GPT-4.5 System Card," 2025.
[35] OpenAI, "GPT-5 System Card," *2025.*
[36] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," *arXiv:2302.13971 ,* 2023.
[37] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull and Esiob, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv:2307.09288,* 2023.
[38] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang and A. Mitra, "The Llama 3 Herd of Models," *arXiv:2407.21783,* 2024.
[39] Meta, "The future of AI: Built with Llama," *Meta,* 2024.
[40] Meta, "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation," *Meta,* 2025.
[41] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, H. Gao, K. Gao, W. Gao, R. Ge, K. Guan, D. Guo, J. Guo, G. Hao, Z. Hao, Y. He and H, "DeepSeek LLM, Scaling Open-Source Language Models with Longtermism," *arXiv:2401.02954,* 2024.
[42] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. Li, F. Luo, Y. Xiong and W. Liang, "DeepSeek-Coder: When the Large Language Model Meets Programming -- The Rise of Code Intelligence," *arXiv:2401.14196,* 2024.
[43] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Luo, G. Hao, G. Chen, G. Li and H. Zhang, "DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model," *arXiv:2405.04434,* 2024.
[44] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo and G. Hao, "DeepSeek-V3 Technical Report," *arXiv:2412.19437,* 2024.
[45] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue and Wan, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," *arXiv:2501.12948,* 2025.
[46] R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen and P, "Gemini: A Family of Highly Capable Multimodal Models," *arXiv:2312.11805 ,* 2023.

[47] P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, S. Mariooryad, Y. Ding, X. Geng, F. Alcober, R. Frostig, M. Omernick, L. Walker and C. Paduraru, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv:2403.05530,* 2024.

[48] G. GeminiTeam, "Gemini 2.0: Our latest, most capable AI model yet," *Google DeepMind,* 2024.

[49] G. GeminiTeam, "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.," *Google DeepMind,* 2025.

[50] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock and Le, "Mistral 7B," *arXiv:2310.06825,* 2023.

[51] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. Bou Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample and L. Lavaud, "Mixtral of Experts," *arXiv:2401.04088,* 2024.

[52] MistralAI, "Au Large," 2024. [Online]. Available: https://mistral.ai/news/mistral-large.

[53] Mistral, "Codestral 25.01," 2025. [Online]. Available: https://mistral.ai/news/codestral-2501.

[54] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain and D. Ganguli, "Constitutional AI: Harmlessness from AI Feedback," *arXiv:2212.08073,* 2022.

[55] Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," *Claude CDN,* 2025.

[56] Anthropic, "Claude Opus 4.1," 2025. [Online]. Available: https://www.anthropic.com/claude/opus.

[57] xAi, "Announcing Grok," 2023. [Online]. Available: https://x.ai/news/grok.

[58] xAi, "Open Release of Grok-1," xAi, 2024. [Online]. Available: https://x.ai/news/grok-os.

[59] xAi, "Grok 4," 2025. [Online]. Available: https://x.ai/news/grok-4.

[60] W. X. Zhao, K. Zhou and J. Li, "A Survey of Large Language Models," *arXiv:2303.18223,* 2023.

[61] D. Korać, B. Damjanović and D. Simić, "A model of digital identity for better information security in e-learning systems," *The Journal of Supercomputing ,* 2022.

[62] C. Pu, J. Seol, N. Park and D. Korac, "Authenticated Key Agreement Protocol for Device-to-Gateway Communication in IoT," in *IEEE Consumer Communications & Networking Conference (IEEE CCNC 2026)*, 2026.

[63] D. Korać, B. Damjanović, D. Simić and C. Pu, "Management of evaluation processes and creation of authentication metrics: Artificial intelligence-based fusion framework," *Information Processing & Management,* 2025.

[64] R. Bommasani, D. A. Hudson and E. Adeli, "On the Opportunities and Risks of Foundation Models," *Stanford CRFM Report,* 2021.

[65] D. Korać, D. Čvokić and D. Simić, "Computational Engineering Approach-Based Modeling of Safety and Security Boundaries: A Review, Novel Model, and Comparison," *Archives of Computational Methods in Engineering ,* 2025.

[66] D. Korać, B. Damjanović, D. Simić and K.-K. R. Choo, "A hybrid XSS attack (HYXSSA) based on fusion approach: Challenges, threats and implications in cybersecurity," *Journal of King Saud University - Computer and Information Sciences,* 2025.

[67] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed., Pearson, 2021.

[68] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.

Received: November 7, 2025
Accepted: November 12, 2025

About the authors

Boris Damjanović is an Associate Professor at the Faculty of Information Technology, Pan-European University Apeiron in Banja Luka. Contact: boris.s.damjanovic@apeiron-edu.eu Areas of interest: data protection in computer systems, artificial intelligence, programming languages, and the application of information technologies.

Dragan Korać is an Associate Professor in the Department of Computer Science, Faculty of Natural Sciences and Mathematics, University of Banja Luka, Bosnia and Herzegovina. He received his Ph.D. in Computer Science and Informatics in 2018 from the Faculty of Organizational Sciences, University of Belgrade. His main research interests include cybersecurity, information security, and mobile computing.

Negovan Stamenković received the B.Sc. degree in Electronics and Telecommunications from the Faculty of Technical Sciences, Kosovska Mitrovica, Serbia, in 2006, and the Ph.D. degree in Electronic Engineering from the Faculty of Electronic Engineering, University of Nis, Serbia, in 2011. He is currently a ˇ Full Professor at the Apeiron, University in BanjaLuka, BiH. His major research interests include digital signal processing, computer engineering, and modular arithmetic.

For citation

Boris Damjanović, Dragan Korać, Dejan Simić, Negovan Stamenković, An Evolutionary Overview of Large Language Models: From Statistical Methods to the Transformer Era, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:145-153, (UDC: 811.163.41`282.4:004.37), (DOI: 10.7251/JIT2502145D), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

FINANCIAL SUSTAINABILITY OF LEARNING PLATFORMS – CASE STUDY OF AN E- LEARNING PROJECT

Sanja Dalton[1], Jefto Džino[2]
[1] *Assistant professor, Belgrade, Serbia, sanja.dalton@metropolitan.ac.rs, 0009 0004 2163 232X*
[2] *Assistant professor, Banja Luka, Bosnia and Herzegovina, jefto.m.dzino@apeiron-edu.eu, 0009 0007 8913 752X*

Preliminary communication
https://doi.org/10.7251/JIT2502154D UDC: 37.018.43:004.738.5

Abstract: This paper analyzes the potential for investment in the implementation of a multifunctional, sophisticated learning platform through a qualitative research approach based on primary and secondary data collected via surveys and semi-structured interviews with high school students, university students, and employees in need of personalized professional development. The analysis and synthesis of the collected data indicate that the commercialization of the innovative platform is feasible.

Keywords: sophisticated platform; investment; platform-based business; personalized learning

INTRODUCTION

Advanced educational technologies are increasingly centered on personalized learning, tailoring content, pace, and teaching methods to meet the specific needs of individual users. Personalized learning platforms serve as dynamic digital tools that utilize user data to deliver optimal educational experiences across diverse target groups. These platforms have proven to be powerful instruments for education across all age groups. Their primary value lies in their capacity to accommodate different learning styles, requirements, and user preferences, thereby improving educational quality, boosting motivation, and enhancing learning outcomes.

Platform-Based Business

Here, we will consider only some of the references in the literature related to platform-based business, also known as a business model defined as a "system consisting of components, the relationships among those components, and the dynamics within the system." [1].

A business model, on one hand, identifies the key participants in a transaction, outlines the value proposition for each party, and describes how the central organization interacts with its partners [7]. On the other hand, it clarifies how value is created, delivered, monetized, and distributed among all stakeholders involved [25]. Several scholars emphasize that, within platform-based business models, it is particularly important to define how value is generated, monetized, and shared between the participants [13]. This includes recognizing the interdependence among activities, which reveals the structural relationships connecting the platform provider and its users [8]. The expansion of digital platforms has significantly reshaped how creative content is produced, shared, and consumed, creating new avenues for creators to monetize their work. Yet, despite the explosive growth of digital content, identifying the most effective monetization strategies remains a persistent challenge [20].

The rapid evolution of digital technologies has fundamentally altered the processes of creating, distributing, and consuming creative content [14]. Social media platforms, for instance, have democratized content creation by enabling individuals to reach global audiences without relying on traditional intermediaries such as publishers or broadcasters [26]. This transformation has allowed creators to profit directly from their work, bypassing conventional industry structures. In the literature, multi-sided

platforms are commonly portrayed as intermediaries or hubs that facilitate value exchange among two or more distinct user groups—often producers and consumers [2][3]. Some researchers define these platforms as networks that link different categories of users to enable transactions [9], while others view them as "interfaces mediating exchanges between multiple parties" [20]. Monetization models vary greatly across platforms such as YouTube, TikTok, Instagram, and Patreon. For example, YouTube's dependence on advertising revenue contrasts with TikTok's focus on brand collaborations and Instagram's influencer-driven approach. Similarly, Patreon's [20] subscription-based model provides creators with a steadier and more direct income stream compared to the variable nature of ad-based earnings.

Monetization strategies are now central to the success of digital creators, with diverse models emerging to address both creator and audience needs [21]. Among the most prevalent methods are advertising revenue, brand partnerships, subscription services, and crowdfunding [17]. Many creators combine multiple revenue streams, using the broad array of monetization tools available on these platforms to enhance their overall earning potential.

Personalized Educational Platforms

Personalized learning methods can help meet individual needs and goals. Personalized learning can be an effective approach that increases motivation, engagement, and understanding [38], maximizing student satisfaction, as well as the efficiency and effectiveness of learning [37]. Some authors argue that the educational community is interested in establishing personalized learning systems that adapt pedagogy, curriculum, and learning environments to meet students' needs and preferences [27]. Others support this claim by noting that a clearly defined concept of personalized learning still does not exist; instead, it is used as an umbrella term for educational strategies that aim to address the individual abilities, knowledge, and needs of each learner [35].

Personalized learning is an educational strategy that adapts instruction to learners' interests, abilities, or needs and usually implies that learners have a certain degree of voice and choice (i.e., autonomy) in that adaptation. Schools, universities, and corporate environments today possess the technological capacity to personalize learning according to the unique needs of each learner. Technology provides numerous options for students and educators to explore new approaches to personalized learning.

Certain authors have proposed a framework for conceptualizing the dimensions of personalized learning in practice [28]. They suggested that personalization of instruction can be achieved by adapting the time, place, pace, and/or path of learning. Other authors have added a fifth dimension to this framework - learning goals [10]. Shemshack [4] and colleagues suggested that a unique, evolutionary approach to personalized learning should encompass four main components: learner profiles, prior learner knowledge, personalized learning pathways, and flexible, self-paced learning environments formed based on dynamic learning analytics. Learning environments that incorporate these various dimensions and components can empower learners to take responsibility for their own learning and strengthen their confidence in achieving learning success.

Research that relies on deconstructions of the concept of personalized learning explains that, although various definitions of personalized learning describe the adaptation of instruction based on students' background, needs, abilities, or interests, descriptions of personalized learning should include the following:

(a) what is personalized – learning objectives, assessments, or educational activities;

(b) how it is personalized – through goals, time, place, pace, and/or learning path;

(c) who or what carries out the personalization – teacher, learner, or adaptive learning system;

(d) on what basis personalization occurs – performance data, activity data, or learner profile data [23].

Other research indicates that further investigation is needed into the outcomes of personalized learning initiatives and the expectation that technology will fulfill its transformative potential in enabling tailored, individualized education [5][24].

Profitability of Platform-Based Business

The continuous evolution of digital platforms has fundamentally reshaped how creative content is produced, distributed, and consumed, creating new avenues for creators to monetize their work. Yet, despite the rapid expansion of the digital content landscape,

identifying the most effective commercialization strategies remains a persistent challenge [20].

Monetization has become a central factor in the success of digital creators, prompting the emergence of diverse models tailored to the needs of both creators and audiences [21]. Among the most widely adopted approaches are advertising revenue, brand collaborations, subscription-based services, and crowdfunding [18]. Many creators employ a mix of these income sources simultaneously, utilizing the variety of tools provided by digital platforms to optimize their earning potential.

Several authors emphasize that a combination of audience engagement, consistent content creation, and platform-specific strategies - such as ad placements, subscription systems, or merchandise sales - plays a vital role in achieving effective monetization. Furthermore, creators who diversify their income streams tend to demonstrate greater financial resilience and long-term sustainability [20].

Despite the surge in online content production, there remains limited understanding of how sustainable these monetization models are over time. While short-term profitability can be achieved, questions persist about how creators can establish durable careers amid continually evolving algorithms, shifting audience interests, and dynamic market trends [39].

Although the growing body of research explores content creation and monetization, the determinants of long-term financial success are still insufficiently examined [22]. While previous studies have investigated individual monetization methods, little is known about how these strategies interact across different platforms [15]. Many creators employ multiple approaches simultaneously, but the ways in which these diverse income streams influence overall financial stability remain underexplored.

Additionally, there is a lack of research examining how different content types - such as educational, entertainment, or lifestyle - affect monetization outcomes [29]. Certain genres may align more effectively with specific monetization models (for instance, tutorial-based content with subscription services), but these relationships have yet to be systematically analyzed. A more nuanced, genre-oriented approach is needed to determine which content categories achieve the greatest success across various platforms [30].

Given the rapid technological advancements shaping digital ecosystems, creators must continuously adapt to emerging formats and evolving user behaviors - factors that significantly influence their ability to sustain revenue generation [33].

Moreover, academic attention has predominantly focused on large-scale influencers, while the financial viability of small and mid-level creators remains underrepresented in research [40]. The unique challenges faced by emerging creators - particularly those lacking institutional backing or resources - are still poorly understood.

Addressing these knowledge gaps in content monetization is crucial for both creators and the platforms that host them [11]. By exploring the combined effects of various monetization mechanisms, creators can better navigate the complexities of the digital creative economy [6]. Insights from such studies could enable the optimization of revenue models and foster long-term economic sustainability.

Furthermore, a deeper understanding of how content types and genres perform within different monetization frameworks could help creators design more targeted, data-driven strategies [34]. This knowledge would also assist digital platforms in developing features that prioritize creators' long-term growth and stability, rather than short-term profit maximization.

In today's digital economy, fueled by the constant consumption of content, monetization has become an indispensable component of creative work. For many, content creation has evolved from a passion into a viable profession. Consequently, exploring alternative income sources beyond traditional advertising - which, while common, often lacks reliability and sufficiency - has become essential for achieving sustainable growth in the creative sector [32].

METHODS AND MATERIALS

This study employs a qualitative research design, focusing on multifunctional platform-based business and, accordingly, the various monetization strategies that content creators can use on digital platforms [16]. The research examines the potential for successfully monetizing work on the platform and investigates the factors that influence creators' financial success [36]. An approach of analysis and synthesis is used for the projected financial analysis of platform monetization.

Sample and Instruments

The sample in this study consists of potential platform users - university students from various faculties, high school students from different schools, as well as employees and professionals in need of personalized professional development. A purposive sampling method was employed to assess projected monetization success, such as steady revenue from advertising, sponsorships, or direct user support (Figure 1).

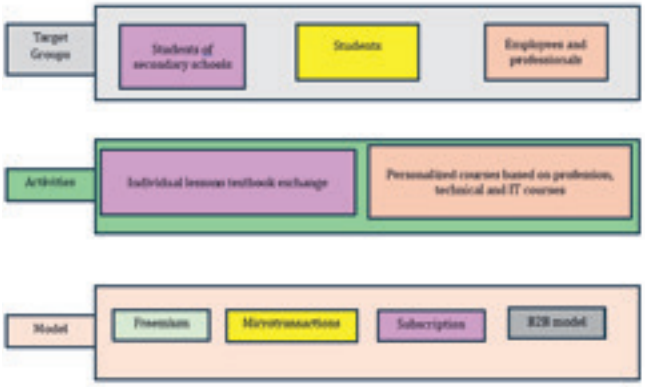


Figure 1. *Target Groups and Strategies of the Innovative Social Network (Source: Author***)**

Data collection was conducted using semi-structured interviews [31]. Interview guides were developed to cover key areas such as creators' monetization strategies, challenges they faced, and financial outcomes. Additionally, secondary data from various platforms and online articles were analyzed to complement the primary data gathered through interviews.

Research Results and Discussion

The data collected for this study include metrics of content demand on the projected platform, such as the exchange of educational and instructional materials, organization of personalized courses, and consultations.

Indicators were projected based on the primary data collected—semi-structured interviews with respondents from target groups—as well as secondary data from platforms where advertisements for these services were posted. Based on the collected data during the research, the projected values of the ratio analysis - total investment, expenses, revenues, and net profit were calculated (Figure 2 and Table 3).

The figure below presents a summary of the main sources of revenue for each platform.

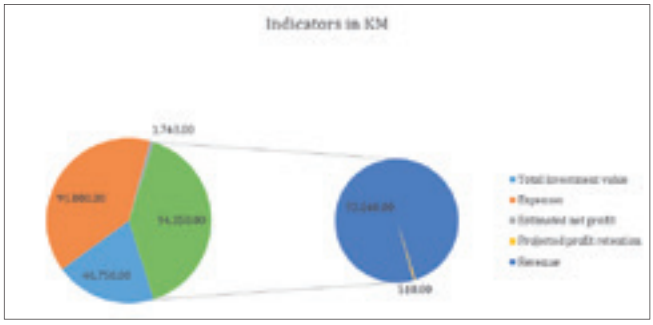


Figure 2. *Projected Investments, Revenues, Expenses, and Net Profit*

Source: Author's processing of research data

The projected financial values and the factors that need to be considered to calculate the required indicators are based on revenue from advertising, subscriptions, and microtransactions, with a significant number of users and interactions contributing to their earnings.

The data reveals an efficiency coefficient through the ratio of achieved effects (results) to expended resources, which is greater than 1. This means that the platform is economically efficient, as each unit of currency invested generates 1.02 units of revenue.

Table 3. *Ratio analysis*

Indicator	Calculation	Value
Efficiency ratio	Total Revenue / Total Expenses	1,02
Accumulation rate	Accumulation/total investment * 100%	1,1 %
PBP	Total investment/cash flow	~ 3,4 years
ROI	Net income /investments * 100%	3,71 %

An analysis of the accumulation rate of 1.1% indicates that a small portion of the profit generated is not spent but reinvested into the further development of the platform. This reinvested amount accounts for 1.1% of the profit set aside for accumulation.

Based on the projected data, the payback period (PBP) is estimated at 3.4 years, representing the time required to recover the initial investment through generated profits or cost savings. This implies that full monetization of the platform is expected to cover the initial investment within this period.

The Return on Investment (ROI) calculation, which assesses the profitability of the investment, shows a positive ROI of 3.71%. This confirms that the investment is economically justified (Table 3).

DISCUSSION

Monetizing a multi-sided platform for personalized learning presents both a significant challenge and a major opportunity for the long-term sustainability and advancement of digital education.

Since personalized learning tailors content, pace, and instructional methods to individual learners, the platform's revenue model must be carefully structured to balance inclusivity, educational quality, and commercial viability.

Launching a platform designed for innovative social learning networks offers the following possibilities:

- Customizable lessons
- Sharing of educational and instructional materials (e.g., scripts, textbooks)
- Personalized consultations
- Profession-oriented courses
- Technical and IT training
- Integration with B2B models

One potential strategy is the **freemium model**, where basic platform functionalities remain free to all users, while premium features - such as mentor access or specialized courses - are offered through subscription plans. This approach allows broad accessibility while ensuring steady revenue from committed users. The freemium concept provides a dual option, enabling users to either use the free version with standard features or upgrade to a paid premium version [19].

Another approach is the **B2B model**, where the platform partners with educational institutions or companies to integrate the system into their internal training and learning programs. In this case, monetization occurs through contracts, training services, or tailored content designed for specific client needs. Numerous studies have explored the impact of the B2B model on platform content revenue [30].

Additional revenue streams include **advertising** and **microtransactions**, allowing users to purchase individual lessons, assessments, consultations, or courses. This model offers increased flexibility and accessibility, particularly for users with limited budgets who prefer selective payments.

Subscription-based models target end users directly, providing creators with more stable and diversified revenue sources that are less susceptible to fluctuations in platform algorithms. Many creators today leverage their loyal audiences by offering exclusive content through subscription services [12].

CONCLUSION

The results of the study, supported by the projected analysis of investments, revenues, and profits (Figure 2 and Table 1), indicate that monetizing an innovative multi-sided platform for personalized learning is feasible, though it demands a well-structured and strategically planned implementation process.

At the outset, adopting a **freemium business model** is recommended. Under this approach, essential platform features - such as the exchange of educational materials and access to general courses - would remain free for all users, while advanced services (including personalized consultations, mentorship programs, and specialized courses) would be monetized through **subscriptions and microtransactions**. This dual strategy promotes inclusivity while generating revenue, a balance particularly valuable in the platform's early growth phase when building a broad user base is a priority.

To maintain consistent income and broaden market presence, it is advisable to **develop B2B collaborations** with educational institutions and corporations. By integrating the platform into their internal training and educational systems, these partnerships can generate contractual revenue streams that enhance long-term financial stability and reinforce institutional cooperation.

Considering the pivotal role of **educational content creators** in digital learning ecosystems [2] [12], implementing a comprehensive incentive and support system for educators is crucial. This system should include transparent revenue-sharing frameworks, increased visibility for high-performing educators, and access to analytical tools that track and interpret user engagement. Such initiatives foster motivation, elevate content quality, and contribute to the professionalization of the overall digital education environment. Moreover, the **user experience must remain central**. Investing in an intuitive interface and responsive customer support can increase retention and reduce barriers to adoption, especially for users with limited digital skills [35].

Following current trends, leveraging **user behavior data** is recommended to improve content personalization. Real-time analysis of user activity can facilitate the timely delivery of relevant courses and resources, enhancing the overall value provided to each user.

Considering the projected ROI of 3.71% and a payback period of 27 months, which falls within a moderately profitable range, it is advisable to **diversify revenue streams**. Combining subscription models, advertising, microtransactions, and possibly crowdfunding for new features reduces reliance on a single source and increases financial resilience.

Ultimately, the successful monetization of a personalized learning platform relies on aligning the business model with both educational objectives and user needs.

The study's limitations include a restricted sample size. The research concludes that the future of monetizing creative content lies in developing **personalized, audience-focused models** and integrating emerging technologies such as artificial intelligence and blockchain, which should be explored in future studies. These insights provide valuable guidance for content creators seeking to optimize monetization strategies in an evolving digital environment.

References

[1] A. Afuah, and C.L. Tucci, *Internet Business Models and Strategies: Tet and Cases*. McGraw-Hill Higher Education, New York (2000).

[2] A. Gawer, and M.A. Cusamano, *Industry platforms and ecosystem innovation*. J. Prod. Innov. Manag. 31(3), 417-433. (2014)

[3] A. Hagiu, and J. Wright, *Marketplace or reseller?* Manag. Sci., 61(1), 184-203. (2015)

[4] A. Schamshack, and J.M. Spector, *A systematic literature review of personalized learning terms.* Smart Learning Environments, 7(33). (2020)

[5] A. Watters, *Teaching Machines: The history of personalized learning.* The MIT Press. (2023

[6] B. Porter, and F. Grippa, *A Platform for AI-Enabled Real-Time Feedback to Promote Digital Collaboration.* Sustainability, 12(24), 10243. (2020) https://doi.org/10.3390/su122410243

[7] C. Baden-Fuller, and V. Mangematin, *Business Models: a challenging agenda*. Strateg. Organ. 11(4), 418-427. (2013)

[8] C. Baden-Fuller, and S. Heafliger, *Business models and technological innovations*. Long. Range Plan., 46(6), 419-426. (2013)

[9] C. Cennamo, and J. Santaló, How to avoid platform traps. MIT Sloan Manag. Rev. 57(1), 12-15. (2015)

[10] C.R. Graham, J. Borup, C.R. Short, and L. Acchambault, *K-12 blended teaching: A guide to personalized learning and online integration*. Provo, UT: Ed Tech Books.org. (2019) http//edtechbooks.org/K12blended

[11] D. Tien Bui, N.-D. Hoang, F. Martínez-Álvarez, P.-T.T. Ngo, P.V. Hoa, T.D. Pham, P. Samui, and R. Costache, *A novel deep learning neural network approach for predicting flash flood susceptibility: A case study at a high frequency tropical storm area.* Science of The Total Environment, 701, 134413. (2020) https://doi.org/10.1016/j.scitotenv.2019.134413

[12] Debutify, *How Content Creators Are Making A Living In The Creator Economy.* Retrieved From https://Debutify.Com/Blog/What-Is-Creator-Economy (2021)

[13] D.J. Teece, *Business models, business strategy and innovation.* Long. Range. Plan. 43(2-3), 172-194. (2010).

[14] D. Mourtzis, J. Angelopoulos, and N. Panopoulos, *A survey of digital B2B platforms and marketplace for purchasing industrial product service systems: A Conceptual Framework*. Procedia CIIRP, 97, 331-336. (2021) https://doi.org/10.1016/j.procir.2020.05.246

[15] E. K. N. Da Silva, V. B. Dos Santos, I. S. Resque, C. A. Neves, S. G. C. Moreira, M. D. O. K. Franco, and W. T. Suarez, *A fluorescence digital image-based method using a 3D-printed platform and a UV-LED chamber made of polyacid lactic for quinine quantification in beverages.* Microchemical Journal, 157, 104986. (2020) https://doi.org/10.1016/j.microc.2020.104986

[16] F. Wu, Y. Wang, J. Y. S. Leung, W. Huang, J. Zeng, Y. Tang, J. Chen, A. Shi, X. Yu, X. Xu, H. Zhang, and L. Cao, *Accumulation of microplastics in typical commercial aquatic species: A case study at a productive aquaculture site in China.* Science of The Total Environment, 708, 135432. (2020)

[17] H. Baltas, M. Sirin, E. Gokbayrak, and A.E. Ozcelik, *A case study on pollution and a human health risk assessment of heavy metals in agricultural soils around sinop province* Chemosphere 241, 125015. (2020) https://doi.org/10.1016/j.chemisphere.2019.125015

[18] H. Baltas, M. Sirin, E. Gökbayrak, and A. E. Ozcelik, *A case study on pollution and a human health risk assessment of heavy metals in agricultural soils around Sinop province, Turkey*. Chemosphere, 241, 125015. (2020) https://doi.org/10.1016/j.chemosphere.2019.125015

[19] I.V. Osipov, E. Nikulchev, D. Plokhou, A.A. Volinsky, (2015). *Study of Monetization as a Way of Motivating Freemium Service Users.* Contemporary Engineering Sciences, 8(20), pp. 911-918, (2015) DOI: 10.12988/ces2015.57212

[20] K. Kadeni, E. Santoso, and W. Jing, *Creative Content Monetization: Case Studies on Digital Platforms*. Journal of Social Entrepreneurship and Creative Technology 2(2), pp. 81-91. (2025) DOI: 10.70177/jsect.vxix.xxx

[21] K.H. Cheng, and C.C. Tsai, *A case study of immersive virtual field trips in an elementary classroom: Students' learning experience and teacher-student interaction behaviors*. Computers & Education, 140, 103600. (2019) https://doi.org/10.1016/j.compedu.2019.103600

[22] K. Kannan, and N. Arunachalam, *A Digital Twin for Grinding Wheel: An Information Sharing Platform for Sustainable Grinding Process.* Journal of Manufacturing Science and Engineering, 141(2), 021015. (2019) https://doi.org/10.1115/1.4042076

[23] L.C. Short, and A. Shemshack, *Personalized Learning.* EDTECHNICA. DOI: 10.59668/371.11067

[24] L. Zhang, J.D. Basham, and S. Yng, *Understanding the implementation of personalized learning: A Research Synthesis.* Educational Research Review, 31 (100339). (2020) https://doi,org/10.1016/J.endurev.2020.100339

[25] M.W. Johnson, C.M. Christensen, and H. Kagermann, *Reinventing your business ecosystems: evidence from application software developers in the iOS and Android smartphone ecosystems.* Organ. Sci. 28(3), 531-551. (2008).

[26] M. Ardolino, N. Saccani, F. Adrodegari, and M. Perona, *A Business Model Framework to Characterize Digital Mul-*

tisided Platforms. Journal of Open Innovation: Technology, Market and Complexity, 6(1), 10. (2020) https://doi.org/10.3390/joitmc6010010

[27] M. Niknam, and P. Thulasiraman, *LPR: a bio-inspired intelligent learning path recommendation system on meaningful learning theory.* Education and Information Technologies. (2020) https://doi.org/10.1007/s10639-020-10133-3

[28] M.B. Horn, and H. Staker, *Blended: Using disruptive innovation to improve schools*. Jossey-Bass. (2014)

[29] M. Sakamiya, Y. Fang, X. Mo, J. Shen, and T. Zhang, *A heart-on-a-chip platform for online monitoring of ontractile behavior via digital image processing and piezoelectric sensing technique*. Medical Engineering & Physics, 75, 36–44. (2020) https://doi.org/10.1016/j.medengphy.2019.10.001

[30] M. Kohtamäki, V. Parida, P.C. Patel, and H. Gabauer, *The relationship between digitalization and servitization. The role of servitization in capturing the financial potential of digitalization.* Technological Forecasting and Social Change, 151, 119804. (2020)

[31] N.P. Harvey Arce, and A. M. Cuadros Valdivia, *Adapting Competitiveness and Gamification to a Digital Platform for Foreign Language Learning.* International Journal of Emerging Technologies in Learning (iJET), 15(20), 194. (2020) https://doi.org/10.3991/ijet.v15i20.16135

[32] O. Adewumni, *Monetization Strategies For Contetnt Creators.* Iosr Journal Of Economics And Finance (Iosr-Jef), 15(6), Ser. 5, pp. 57-66. E-Issn: 2321-5933, P-Issn: 2321-5925. (2024) www.Iosrjournals.Org https://doi.org/10.1016/j.scitotenv.2019.135432

[33] P. Samui, J. Mondal, and S. Khajanchi, *A mathematical model for COVID-19 transmission dynamics with a case study of India.* Chaos, Solitons & Fractals, 140, 110173. (2020) https://doi.org/10.1016/j.chaos.2020.110173

[34] Q. Wang, and M. Su, *A preliminary assessment of the impact of COVID-19 on environment –A case study of China.* Science of The Total Environment, 728, 138915. (2020) https://doi.org/10.1016/j.scitotenv.2020.138915

[35] R. Schmid, and D. Petko, *Does the use of educational technology in personalized learning environments correlate with self-reported digital skills and believes of secondary-school students?* Computers & Education, 136 (March), 75-86. (2019) https://doi.org/10.1016/j.compedu.2019.03.006

[36] R. Xiao, D. Guo, A. Ali, S. Mi, T. Liu, C. Ren, R. Li, and Z. Zhang, *Accumulation, ecological-health risks assessment, and source apportionment of heavy metals in paddy soils: A case study in Hanzhong, Shaanxi, China.* Environmental Pollution, 248, 349–357. (2019) https://doi.org/10.1016/j.envpol.2019.02.045

[37] S. Gómez, S.P. Zerva, D.G. Sampson, and R. Fabreget, *Context-aware adaptive and personalized mobile learning delivery supported by UoLmP*. Journal of King Sand University – Computer and Information Sciences, 26(1), 47-61. (2014) https://doi.org/10106/j.jksuci.2013.10.008

[38] T. Pontual Falcão, F.M.A. e Peres, D.C. Sales de Morais, and G. da Silva Oliveira, *Participatory methodologies to promote student engagement in the development of educational digital games*. Computers & Education, 116, 161-175. (2018) https://doi.org/10.1016/j.compedu.2017.09.006

[39] W. Feng, Q. Zhang, H. Ji, R. Wang, N. Zhou, Q. Ye, B. Hao, Y. Li, D. Luo, and S. S. Y. Lau, *A review of net zero energy buildings in hot and humid climates: Experience learned from 34 case study buildings.* Renewable and Sustainable Energy Reviews, 114, 109303. (2019) https://doi.org/10.1016/j.rser.2019.109303

[40] W. Gao, P. Veeresha, H. M. Baskonus, D. G. Prakasha, and P. Kumar, *A new study of unreported cases of 2019-nCOV epidemic outbreaks.* Chaos, Solitons & Fractals, 138, 109929. (2020) https://doi.org/10.1016/j.chaos.2020.109929

[41] X. Xu, Q. Zhang, J. Song, Q. Ruan, W. Ruan, Y. Chen, J. Yang, X. Zhang, Y. Song, Z. Zhu, and C. Yang, *A Highly Sensitive, Accurate, and Automated Single-Cell RNA Sequencing Platform with Digital Microfluidics*. Analytical Chemistry, 92(12), 8599–8606. (2020) https://doi.org/10.1021/acs.analchem.0c01613

Received: October 31, 2025
Accepted: November 18, 2025

ABOUT THE AUTHORS

Sanja Dalton is an Associate Professor professor at Metropolitan University in Belgrade. She earned her PhD at the Faculty of Organizational Sciences, University of Belgrade, specializing in Innovation and Technological Development. Her areas of expertise and research include digital innovation and business process management.

Jefto Džino was born in 1966 in Centar-Sarajevo. He is an Assistant Professor at the Faculty of Information Technologies at the Pan-European University APEIRON in Banja Luka. He completed his master's studies in 2011 at Metropolitan University Belgrade, Faculty of Information Technologies, and earned his PhD in 2021 at the Faculty of Information Technologies, Pan-European University APEIRON. His main research interests include Information Systems, Business Intelligence, Artificial Intelligence, and e-Government.

FOR CITATION

Sanja Dalton, Jefto Džino, Financial Sustainability of Learning Platforms – Case Study of an E- Learning Project, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:154-160, (UDC: 37.018.43:004.738.5), (DOI: 10.7251/JIT2502154D), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

Model to improve distance learning system LOOMEN

Karlo Čuković - Tkalčec
Koprivnički Bregi Elementary School, Koprivnica, Croatia, kcukovic@gmail.com

Professional paper
https://doi.org/10.7251/JIT2502161T UDC: 004.738.5:37.018.43

Abstract: This research study focuses on analyzing the functionalities and potential improvements of the Loomen platform, the most widely used LMS system in Croatian education. The aim of the study was to identify the strengths and weaknesses of the platform and propose solutions that could enhance user experience and improve pedagogical outcomes. The research was conducted during a teaching internship at a high school, where hands-on experience with daily system use was gathered, and through a survey administered to students and teachers. The survey collected both quantitative and qualitative data on usage frequency, technical difficulties, satisfaction with functionalities, and suggestions for improvement. The results indicate that students and teachers value the ability to access teaching materials and submit assignments, but they highlight issues with the user interface, mobile version, and communication tools. Students perceive communication via forums and messaging as outdated and insufficiently engaging, which reduces interaction in the learning process. Based on the collected data, a list of functionalities and identified shortcomings was compiled, followed by proposals for improvement: interface redesign, optimization for mobile devices, introduction of a self-assessment module, and richer multimedia content. The paper concludes by emphasizing that the proposed measures have the potential to increase student motivation and satisfaction, as well as to improve the quality of distance learning.

Keywords: e-learning, LMS, Loomen, digital education

Introduction

Education in recent decades has been undergoing the most dynamic changes in its history. The development of digital technologies, high-speed internet, mobile devices, and artificial intelligence has introduced completely new ways of teaching and learning, pushing the boundaries of the traditional classroom and paving the way for the concept of lifelong learning. Today, knowledge is no longer transmitted exclusively through direct interaction between students and teachers, but also through digital platforms and educational systems that enable learning "anytime and anywhere" [2].

This process of digital transformation in education coincides with the development of so-called LMS (Learning Management System) platforms – information systems designed for organizing, managing, and evaluating learning. LMS systems allow teachers to create, publish, and monitor educational activities, while providing students with a centralized environment for accessing materials, assignments, and feedback. According to [2], LMS has become the foundation of modern e-learning, as it provides a unique meeting point between pedagogy, technology, and the management of the educational process.

In the Croatian education system, the most widely used LMS platform is Loomen, developed and maintained by CARNET. It is based on the open-source system Moodle, which is globally recognized as a flexible and adaptable open-source platform. Loomen was conceived as a national digital educational tool that enables students and teachers to easily exchange teaching materials, participate in online tests, communicate through forums, and store their learning results. During the COVID-19 pandemic, Loomen, together with Microsoft Teams, became a key tool for ensuring the continuity of education in Croatia—from primary and secondary schools to higher education institutions. This situation rapidly accelerated the process of digital transformation but also revealed a number of limitations of the existing digital tools [3].

An analysis conducted during the preparation of this paper showed that, despite its overall functional-

ity and widespread use, Loomen does not fully exploit its pedagogical potential in practice. Although it allows for the digital organization of teaching, the pedagogical and motivational components of the system often remain underdeveloped. The mere availability of tools does not automatically ensure their effectiveness— as [6] emphasizes, successful e-learning requires a synergy between technical performance and pedagogical design, with technology serving educational goals rather than dictating them.

Experiences of students and teachers gathered through a survey conducted in a high school setting further confirm this assumption. The participants most frequently highlighted several issues: an unclear user interface, slow system performance, poor optimization for mobile devices, and limited possibilities for interactive communication. Students, for instance, noted that forum communication feels "outdated" and that they prefer using external applications such as WhatsApp or Discord for collaboration. They also emphasized the need for greater content diversity—more video lessons, multimedia presentations, and interactive exercises instead of predominantly static Word and PowerPoint documents. Such findings confirm results from other studies suggesting that today's students, as so-called digital natives, prefer dynamic, visually engaging, and interactive content [5].

Teachers, on the other hand, expressed satisfaction with functions that simplify administration and grading but pointed out that the process of creating digital materials is too time-consuming, and that the tools for monitoring student activity and providing automated feedback are limited. Their comments indicate a need for a smarter system of tracking and self-assessment that would allow both students and teachers a clearer insight into progress and understanding of the material.

All of this indicates that Loomen has reached a critical stage in its development: it is technically stable and widely implemented but lacks the elements that would make it pedagogically more effective and motivationally engaging. In the context of modern digital pedagogy—where the emphasis is on active learning, self-regulation, and collaboration—such aspects are crucial for the success of digital education [6].

The aim of this paper is to analyze the current state of Loomen, identify its key strengths and weaknesses, and, based on the obtained data and theoretical framework, propose a model for improving the platform. The model will cover four main areas:

1. Improving the user interface and navigation
2. Introducing modules for self-assessment and reflection
3. Integrating multimedia and interactive content
4. Ensuring technical stability and better mobile adaptation

The paper does not deal with the technical implementation of the system but rather provides a **pedagogical-technological framework** that illustrates how the platform should be designed to support modern forms of learning. In this sense, Loomen is viewed as a living system that evolves along with the needs of its users and digital trends.

In conclusion, this paper starts from the premise that technology alone does not guarantee quality learning—it does so only when it serves an educational purpose. Therefore, improving Loomen does not merely imply technical modernization but primarily the alignment of technological solutions with pedagogical goals and the real needs of students and teachers. In this way, the main objective of the paper is achieved: to develop a sustainable model of digital education that is functional, motivating, and centered on the learner as an active participant.

RESEARCH METODOLOGY

To obtain a comprehensive and objective insight into the functionalities and capabilities of the Loomen platform, this study applied a combined methodological approach that includes an analysis of relevant literature, an evaluation of the system's functionalities through practical use of the platform, a comparative analysis with other LMS solutions, and the collection of feedback from actual users. Such an approach enables the integration of quantitative and qualitative indicators and provides a holistic understanding of the system's technical, pedagogical, and user-oriented aspects.

Research Framework and Objectives

The main objective of this research was to identify the key strengths and weaknesses of the **Loomen** platform from the perspective of end users — students and teachers — and to propose recommendations for its

improvement in both technical and pedagogical terms. The specific objectives were:

1. To analyze the existing functionalities of Loomen and their pedagogical relevance
2. To determine which usability issues most affect learning effectiveness
3. To examine the user experience of students and teachers
4. To propose a set of improvements aligned with the theoretical and practical needs of e-learning

The methodology was designed in accordance with the recommendations of [2], who distinguishes three levels of scientific research: theoretical (literature review), empirical (data collection and analysis), and applied (implementation of results in the form of improvement proposals).

Platform Analysis and Experiential Testing

The first stage of the research involved a detailed examination of the Loomen platform's functionalities. For this purpose, a test user account was created with two levels of access — teacher and student roles — which enabled a comprehensive insight into the system from both perspectives. The testing process covered the following key functions:

1. Creation and distribution of teaching materials
2. Design and administration of online tests
3. Assignment submission and grading
4. Monitoring of student progress
5. Communication through forums and internal messages
6. Integration of external content (images, PDFs, videos, H5P activities, etc.)

Special attention was given to the clarity of the user interface, system responsiveness, and mobile adaptability, as these elements have often been identified in previous studies as key factors influencing user satisfaction [4]. Figure 1 shows the home page of the test course created within the subject *Informatics*.

User Data Collection

Another important segment of the research involved collecting data from actual users — high school students and teachers. A survey was created using Google Forms and distributed digitally to ensure accessibility and respondent anonymity. The survey included 80 students and 4 teachers who regularly use the Loomen platform in their teaching activities.

The questionnaire consisted of 18 questions divided into three sections:

1. Frequency and manner of platform use (e.g., how often users access Loomen, from which devices, and for what purposes),
2. User experience and satisfaction level (evaluation of functionalities such as tests, forums, and clarity of learning materials),
3. Suggestions and comments for improvement (open-ended questions).

The responses were analyzed using a combination of descriptive statistics (expressed in percentages and frequencies) and thematic analysis for qualitative responses. This approach provided insight not only into what users do on the platform, but also into

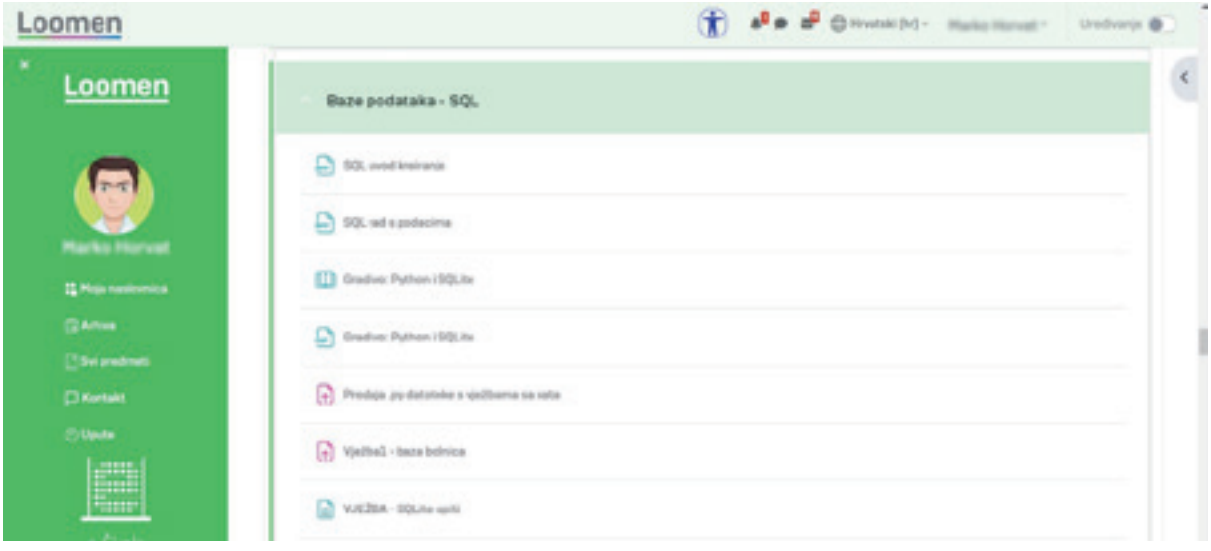


Figure 1- *Homepage of the Loomen Course: Informatics*

Figure 2- *High School Students Participating in the Survey on the Functionality of Loomen*

how they experience it.

Figure 2 shows high school students who participated in the survey on the use of Loomen.

In addition to collecting user feedback, a comparative analysis was conducted with several other popular LMS platforms, including Moodle (on which Loomen is based), Microsoft Teams, and Google Classroom. The analysis involved a review of their functionalities, methods of integration with other tools, and the flexibility of adapting to teachers' needs. This comparison proved valuable in identifying both the advantages that Loomen already possesses and the functionalities that are better implemented in other systems and could potentially be incorporated into Loomen in the future.

The collected data were then categorized into three main groups:

- Technical features (system stability, loading speed, mobile support),
- Pedagogical features (progress tracking, student motivation, interactivity), and
- User experience (interface clarity, ease of use, communication options).

This categorization helped clearly distinguish issues arising from the system's technical limitations from those related to didactic design and content organization.

Based on the collected and analyzed data, a list of functionalities currently available in Loomen was

Table 1-Current *Functionalities of the Loomen Platform*

Functionality	Description
Access to Teaching Materials	Students can access documents (PDF, Word, PPT) and multimedia files published by teachers.
Assignment Submission	Enables uploading and submitting assignments within specified deadlines.
Tests and Quizzes	Allows the creation and completion of tests and quizzes with automatic grading.
Forum and Internal Messages	Communication between students and teachers through discussion forums and private messages.
Grades and Feedback	Provides an overview of grades and teacher comments for each assignment or test.
Progress Tracking	Displays student activity, lesson completion, and task progress.

Table 2- *Identified Weaknesses and Proposed Improvements*

Identified Weakness	Proposed Improvement
Unclear user interface	Redesign the interface with a focus on simpler navigation and a more modern appearance.
Difficult login and occasional technical issues	Improve system stability and enable faster login, especially during periods of high demand.
Outdated communication tools (forum, messages)	Integrate a modern chat module and enable video communication features.
Lack of multimedia content	Increase the use of video lessons, interactive quizzes, and presentations.
Weak mobile version	Ensure full optimization for mobile devices and tablets.
Insufficient options for self-assessment	Introduce a module for self-check and content revision.

created (Table 1), along with a separate list of shortcomings and potential improvements (Table 2).

Comparative Analysis with Other LMS Platforms

To place Loomen within a broader context, a comparative analysis was conducted with three of the most widely used LMS solutions:

- Moodle
- Microsoft Teams
- Google Classroom

The analysis included an evaluation of core functionalities, user adaptability, multimedia integration, and options for tracking student progress. The results showed that Moodle, as the original platform on which Loomen is based, offers the greatest flexibility in configuration, while Teams and Classroom provide simpler but less customizable systems. This confirmed that Loomen has a solid technical foundation, yet requires modernization of the user experience to keep up with its competitors.

Although this paper does not describe the technical implementation of proposed solutions, the recommendations serve as guidelines for further development of the platform and may act as a starting point for future researchers and IT specialists working on its improvement.

Such a comprehensive methodological approach ensures that the analysis is not limited to the author's subjective impression but is based on a combination of experiential data, survey results, and comparative analysis. In this way, the findings gain greater credibility and can be considered representative of the broader population of platform users.

Analytical Framework of the Research

All collected data — surveys, testing notes, and comparative analysis (Table 3) — were classified into three main categories that constitute the analytical framework of the research.

This framework made it possible to systematically distinguish between problems arising from techni-

Table 3- *Analytical Framework of the Research*

Category	Key Elements	Purpose of the Analysis
Technical Features	Stability, speed, mobile support, integration	To assess the system's reliability and accessibility
Pedagogical Features	Interactivity, motivation, formative assessment	To determine the system's impact on the learning process
User Experience (UX/UI)	Clarity, intuitiveness, communication	To evaluate user satisfaction and engagement

cal limitations and those resulting from insufficient pedagogical design.

Validation and Limitations of the Research

To ensure credibility, the survey results were compared with available studies on the use of the Moodle and Loomen platforms within the Croatian educational context [3]. Although the sample was limited to one high school, the obtained results provide valuable insights into user experience that can be applied more broadly within the context of secondary education.

The main limitations of the research relate to:

- the small number of respondents in the teacher group,
- the potential subjectivity of survey responses, and
- the absence of technical implementation of the proposed solutions.

Despite these limitations, the combination of empirical data, testing, and comparative analysis makes this methodology a relevant and reliable basis for formulating recommendations for improving the platform.

The research process and its stages are presented in Figure 3.

RESULTS AND DANA ANALYSIS

Based on the conducted analysis of the **Loomen** platform's functionalities, as well as the survey carried out among students and teachers, a comprehensive overview of the current state of the system was obtained. The results are presented in several thematic sections covering available functionalities, the most common issues encountered when using the platform, frequency of use, and user suggestions for improvement. Such a presentation provides a holistic insight not only into the system's technical capabilities but also into how it is perceived and utilized in real educational settings.

Currently Available Functionalities

The Loomen platform enables a wide range of activities that support the teaching process and facilitate communication between students and teachers. The most significant among these include:

- **Access to teaching materials** – Teachers can upload learning content in the form of MS Word documents, PowerPoint presentations, PDF files, images, and links to external sources. The materials are organized by topics and weeks, which allows easier navigation and systematic tracking of the curriculum. For students, this means convenient access to learning materials at any time, without the need for additional tools or platforms.
- **Test creation and administration** – The quiz and test creation feature enables teachers to design various types of questions, including multiple choice, true/false, fill-in-the-blank, matching, and essay questions. The system automatically grades closed-ended questions, while open-ended ones are assessed manually, giving teachers greater flexibility in evaluation. Students particularly appreciate the immediate feedback provided after each test, which encourages revision and self-reflection.
- **Assignment submission and grading** – Digital submission of assignments significantly improves organization and reduces the administrative workload for teachers. Students can submit their work in various formats, and the system records submission times, ensuring transparency. Teachers can add comments and suggestions for each submission, fostering a more effective learning process through constructive feedback.
- **Communication** – Loomen offers a discussion forum for each course unit, the ability to send private messages, and post announcements about upcoming activities. Although these tools are functional, they tend to be underused in practice, as students prefer faster and more modern communication channels. However, the forum still has potential as a tool for structured discussions and exchange of ideas, especially when combined with moderated assign-

Figure 3- *Research Flow Diagram*

ments and group projects.

- **Student progress tracking** – Teachers have access to students' activity data, test results, and the number of visits to course materials. This information helps identify students who are struggling or not actively participating, enabling early detection of learning difficulties.
- **Multimedia integration** – The system supports embedding videos, audio files, external links, and H5P interactive content, allowing the creation of more dynamic and engaging lessons. This makes the learning process more visually appealing and adaptable to different learning styles.

Although these functionalities form a solid foundation for conducting online instruction, their usability and quality of implementation largely depend on the teachers' **digital competencies** and the school's **technical infrastructure**. Some teachers take advantage of advanced features, while others use the platform solely as a repository for teaching materials. This results in considerable variation in the quality of user experience among students across different subjects and teachers.

Frequency of Use and User Satisfaction

The results of the survey conducted among students and teachers indicate that Loomen is becoming an increasingly common tool in the daily educational process, although its frequency of use still varies significantly across subjects and depends largely on the individual teacher's approach. In most cases, the platform is used regularly, but it has not yet been fully integrated as the main learning tool; rather, it primarily serves as a support to traditional teaching.

Teachers most frequently use Loomen for publishing teaching materials, assignments, and tests, while communication tools such as forums and messages are less utilized. Most teachers emphasized that forums require more time to manage and fail to attract students who prefer faster, more modern, and visually dynamic platforms.

User satisfaction reveals two opposing tendencies. On one hand, students highlight the practicality and organization of Loomen—the fact that they can access all learning materials in one place and more easily keep track of their assignments. On the other hand, users report technical difficulties, particularly occasional system slowdowns and the limited functionality of the mobile version. A large number of students stated that they most often access the platform via smartphones but that it "sometimes responds slowly or fails to load all elements correctly."

These findings indicate the need for further technical optimization, especially in terms of mobile accessibility and system stability. At the same time, the high level of satisfaction with the platform's basic functionalities suggests that Loomen already effectively supports digital learning, but certain improvements are required to increase student motivation, interactivity, and overall platform efficiency.

Identified problems

The results of the survey conducted among students and teachers show that Loomen is becoming an increasingly common tool in everyday educational practice; however, the frequency of use varies depending on the subject and the teacher's approach. According to the collected data, 68% of students use the platform at least once a week, 24% several times a week, while only 1% use Loomen on a daily basis. These results suggest that the platform has not yet been fully integrated into the teaching process as the main learning tool, but rather serves primarily as support to traditional instruction.

Teachers most frequently use Loomen for publishing teaching materials, assignments, and tests, while communication tools and forums are less commonly used. In interviews and open-ended survey responses, most teachers emphasized that maintaining the forum requires additional time and that it fails to engage students who are accustomed to faster and more visually appealing platforms.

User satisfaction demonstrates two contrasting tendencies. On one hand, students expressed strong satisfaction with the ability to access all learning materials in one place, noting that Loomen helps them organize their learning and keep track of assignments. On the other hand, both students and teachers reported technical difficulties, including occasional system slowdowns, crashes during peak usage, and problems when accessing the platform via mobile devices. More than half of the students (52%) stated that they most often access Loomen through their smartphones, but that "pages sometimes do not respond properly or load too slowly."

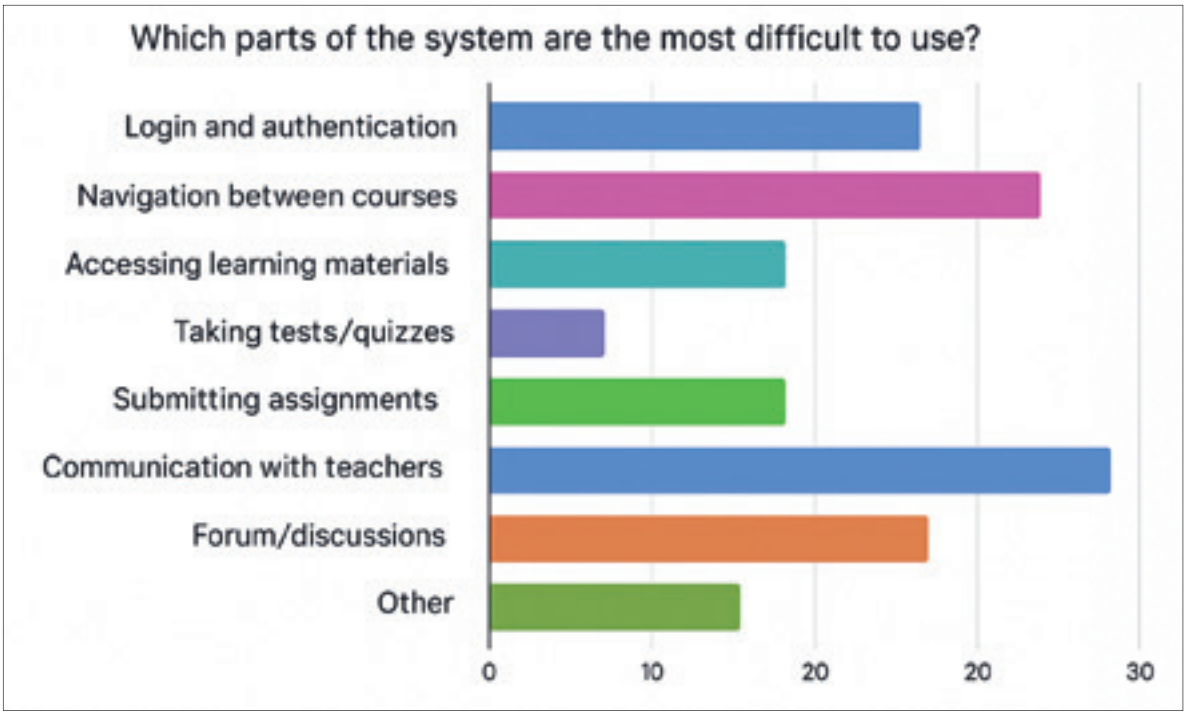


Figure 4- *The most difficult parts of the system for users to use*

These findings clearly indicate the need for greater focus on mobile optimization and improved system stability. At the same time, the level of teacher satisfaction with the available functionalities shows that Loomen generally fulfills its intended purpose but requires further adjustments to enhance student motivation and the overall efficiency of use.

Suggestions for Improvement

Based on the collected results and user feedback, several concrete recommendations for improving the Loomen platform can be identified (Figure 5):

- **Redesign of the user interface** – The platform layout should be simplified, allowing personalization options (e.g., topic filtering, highlighting deadlines, visually marking submitted assignments) and improved navigation. This would enhance clarity and increase student motivation to use the system.
- **Enhancement of the mobile version** – The mobile application should be as functional as the desktop version, with optimized loading speed, better responsiveness, and improved screen adaptation. Students noted that such an upgrade would significantly increase the frequency of platform use.
- **Improved system stability and reliability** – Better server optimization and technical support during peak loads would reduce user frustration and increase trust in the system.
- **Development of a self-assessment module** – Introducing options for quick, anonymous knowledge checks (mini-quizzes, self-assessment after lessons) would promote independent learning and enhance the platform's pedagogical value.
- **Greater use of multimedia content** – Teachers should be encouraged to integrate interactive materials (videos, H5P content, virtual simulations). Training sessions and workshops could help strengthen their digital competencies.
- **Improved communication tools** – Implementing more modern communication features, such as group chats, push notifications, and simplified messaging, would foster greater collaboration and faster information exchange.

These recommendations show that Loomen already possesses all the essential technical prerequisites for effective e-learning, but the system still needs to be further adapted to the habits of today's students and the demands of modern education. Increasing interactivity, simplicity, and technical stability are key steps toward transforming Loomen from

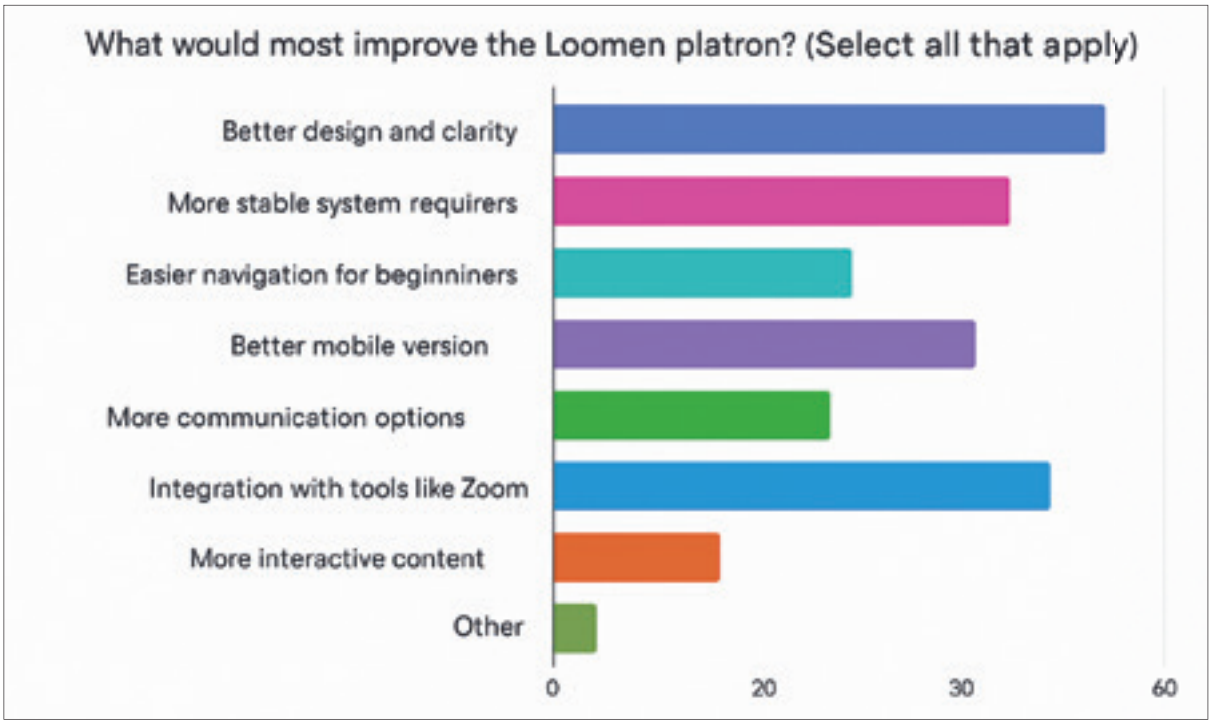


Figure 5- *Suggestions for improving the Loomen platform*

a functional platform into a fully integrated digital learning environment.

Discussion

Analysis of the collected data on the use of the Loomen platform provided a clearer picture of how students and teachers perceive this tool and how they use it in everyday educational practice. Although the results show that most users regularly use Loomen and find it useful, the experiences are mixed—technical reliability and pedagogical effectiveness are not always perceived in the same way. The platform is technically functional, but its pedagogical value depends on how it is implemented, the teachers' digital literacy, and the students' motivation.

During my internship at a secondary school, I had the opportunity to work as a teacher using the Loomen system, preparing teaching materials and monitoring student activities. This experience allowed me to view the research results not only through numbers but also through real classroom situations. I was able to observe firsthand the difference between the theoretical possibilities of the system and the way it is actually used in practice.

The greatest advantage of Loomen remains its accessibility and flexibility. Students can complete assignments outside of the classroom, often late in the evening or on weekends, which is particularly useful for those who learn at their own pace or wish to make up missed lessons. This flexibility has proven to be a key motivational factor—students report that the ability to "learn whenever they want" gives them a sense of control over their own learning. This confirms the view of Garrison and Anderson (2003) that learner autonomy is a fundamental component of successful e-learning.

On the other hand, discussions with teachers revealed that technical reliability is one of the biggest challenges. Teachers often express frustration with occasional login difficulties, slow page loading, or system crashes during exam periods. These issues affect not only the technical delivery of lessons but also the overall perception of the reliability of digital tools. When the system fails at critical moments, user trust diminishes, which can reduce teachers' willingness to adopt digital innovations.

A particularly interesting observation concerns the type and format of learning materials. Students greatly value lessons that include videos, interactive quizzes, images, and practical examples compared to traditional text-based documents. During my teaching experience, I noticed that engagement levels were

significantly higher with multimedia content: students responded faster, participated more actively in discussions, and were more likely to complete their assignments. These findings confirm [6] principles of multimedia learning, which suggest that the combination of text, visuals, and audio enhances cognitive engagement.

Pedagogical analysis also indicates that Loomen provides a solid framework for formative assessment, but its potential is not fully utilized. Teachers rarely use automatic grading and feedback features, even though these tools can encourage continuous learning and self-assessment. Introducing modules for self-evaluation and "mini-tests" without grades could create a more supportive environment where students monitor their progress without fear of making mistakes. Such features contribute to the development of self-regulated learning, which is considered one of the key goals of digital education according to modern pedagogical models.

From a technical perspective, it is clear that Loomen, as Croatia's implementation of Moodle, has a strong and stable open-source architecture. However, technical stability alone is not sufficient without continuous updates and optimization. In practice, most issues occur during periods of intensive use (exam weeks, end of semester), when the system slows down or becomes temporarily unavailable. This points to the need for better resource management and ongoing technical monitoring, as well as for strengthening user support to resolve issues more efficiently.

From a user experience (UX/UI) standpoint, both students and teachers notice that the interface is not sufficiently intuitive or modern. Most users describe Loomen as functional but "outdated" and "overly complex." Compared with tools such as Google Classroom or Microsoft Teams, Loomen appears technically more complex but less visually accessible. Google Classroom stands out for its simplicity and clarity, while Teams provides an integrated environment with instant notifications and chat. Loomen, while maintaining its flexibility, could benefit greatly from adopting a more modern, responsive design with clear icons, personalization options, and shorter navigation paths to key functions.

The pedagogical effectiveness of the platform also depends on how teachers design their activities. The system itself cannot ensure quality learning if it is not used in a didactically meaningful way. Teachers need to be familiar with active and constructivist learning methods, where students are not passive recipients of information but active participants in the learning process. In this regard, CARNET, as the system provider, could develop educational programs to help teachers fully utilize the potential of Loomen—not only as a repository for files but as a digital space for exploration, collaboration, and reflection.

When compared with other LMS solutions such as Moodle, Microsoft Teams, and Edmodo, Loomen shares many functionalities but differs in terms of local adaptation and accessibility. Teams, for example, is strongly integrated with the Office suite and enables synchronous communication, while Moodle (and Loomen) focuses more on asynchronous learning and detailed progress tracking. Edmodo, on the other hand, emphasizes simplicity and social interaction but offers less control. This comparison shows that Loomen occupies a middle ground—it has a strong structure and control but could benefit from greater intuitiveness and aesthetic appeal.

Within the context of the Croatian educational system, Loomen holds additional value because it is free, nationally supported, and integrated with e-Matica and e-Dnevnik, which gives it a significant institutional advantage. However, technological capabilities and pedagogical potential must remain balanced. If students do not perceive the platform as engaging and useful, its technical complexity loses its purpose.

As a future teacher and system user, I believe that integrating a modern design, ensuring greater stability, and diversifying teaching materials could make Loomen more effective. This would not only improve user satisfaction but also foster higher student motivation, better interaction, and more successful knowledge acquisition. E-learning would thus become not merely an emergency alternative but an equally valuable and desirable method of education in the digital age.

CONCLUSION

The conducted research provided a comprehensive insight into the current state of the Loomen platform and the experiences of its users. Based on the collected data, surveys, and personal experience during teaching practice, it became clear that Loomen

has become an indispensable tool in the modern educational system of the Republic of Croatia, especially in the context of distance and hybrid learning models. This confirms the findings of [1], which emphasize the importance of e-learning and LMS platforms in increasing access to educational materials and creating a more inclusive learning environment.

Its core functionalities—access to learning materials, test creation and administration, and communication through forums—have been well received by both students and teachers. The platform largely fulfills its primary purpose: digitally connecting all participants in the educational process and ensuring continuity of teaching regardless of time and place constraints. This is consistent with [5], who highlight that the basic functionalities of LMS platforms are the foundation of pedagogically effective and sustainable e-learning systems.

However, numerous comments and survey results indicate that the system still has significant room for improvement. The main hypothesis of this research—that improving the user interface increases student efficiency and satisfaction—was confirmed through both quantitative data and practical experience. Students and teachers almost unanimously emphasize the need for a simpler, clearer, and more intuitive interface that allows faster navigation and reduces frustration during use. This conclusion aligns with [3], who note that a well-designed interface directly influences learner motivation, engagement, and autonomy.

In the modern educational context, design is not merely an aesthetic but also a pedagogical category—an intuitive interface encourages activity, while a complex or outdated one diminishes concentration and interest.

The secondary hypothesis regarding the importance of self-assessment modules was also confirmed by the research results. Students expressed a strong desire for the ability to independently check their knowledge, which would allow them to track progress without the pressure of formal grading. This would help develop key competencies such as self-regulation, responsibility, and self-motivation. These results are consistent with [2], who emphasize the importance of interactive and self-directed learning in digital environments. Introducing such modules would not only increase the pedagogical value of Loomen but also contribute to a broader goal—the development of students who take an active role in their own education.

The third hypothesis, relating to system stability and mobile optimization, emerged as the most significant user demand. A large number of students reported that technical difficulties—such as slow loading, login problems, and poor mobile performance—significantly hinder their use of the platform. This supports the conclusions of [3], which underline that technical reliability and mobile accessibility are key factors of user experience in any LMS system. In today's context, where most students use smartphones as their primary learning device, optimizing the mobile version is no longer optional but essential.

The analysis conducted in this study also showed that technical solutions, regardless of how advanced they are, are insufficient without pedagogical understanding and user-centered design. An e-learning system must be designed to foster interaction, collaboration, and reflection—three key elements of effective digital pedagogy. In this sense, Loomen holds significant potential, but continued development is needed in three key areas:

- **Technical stability** – system optimization, improved user support, and continuous functionality testing.
- **Pedagogical innovation** – development of teachers' digital competencies and encouragement of interactive and multimedia content.
- **User experience (UX/UI)** – modernization of the interface, visual clarity, and intuitive navigation.

Implementing the proposed improvements would have multiple benefits. On one hand, technical barriers would be reduced and teacher efficiency increased, while on the other, students would gain a more motivating and interactive learning environment. This would create conditions for deeper understanding of learning content, the development of critical thinking, and greater satisfaction with the learning process.

These conclusions extend beyond Loomen itself to the broader understanding of digital transformation in Croatian education. In recent years, especially after the COVID-19 pandemic, it has become evident that digital solutions are no longer an addition but an integral part of the education system. As a national plat-

form, Loomen symbolizes this shift but also carries a responsibility—to continue evolving in line with the needs of new generations of students and teachers.

From a pedagogical perspective, this research confirms that technology alone does not create learning—it occurs through meaningfully designed activities, supportive interfaces, and stimulating environments. Therefore, future research should focus not only on the system's technical development but also on monitoring the effects of new functionalities on actual learning outcomes. It is recommended to conduct pilot projects that include a modernized interface, enhanced communication tools, and self-assessment modules, accompanied by systematic tracking of their impact on student motivation and achievement over time.

Loomen already plays a key role in Croatia's educational landscape. Its future development should be guided by the principles of simplicity, interactivity, and accessibility, so that the platform becomes not only a functional tool but also an inspiring digital environment that motivates students and supports teachers.

Ultimately, it can be concluded that Loomen is not merely a technical system but a **pedagogical bridge** between traditional and digital schooling. If the recommended improvements are implemented, the platform could become a model of effective e-learning—one that goes beyond necessity and establishes itself as a standard of quality education in the 21st century.

REFERENCES

[1] Stanković Ž. i Petrović, M. (2016) *E-učenje*, Banja Luka: Panevropski univerzitet Apeiron

[2] Ćukušić, M. i Jadrić, M. (2012) *E-učenje: koncept i primjena*, Zagreb: Školska knjiga

[3] Hoić-Božić, N. i Holenko, M. (2021) *Uvod u e-učenje: obrazovni izazovi digitalnog doba*, Rijeka: Sveučilište u Rijeci, Odjel za informatiku

[4] Zelenika, R. (2000) *Metodologija i tehnologija izrade znanstvenog i stručnog djela*, Rijeka: Ekonomski fakultet Sveučilišta u Rijeci.

[5] Bjekić, D., Krneta, R., & Milošević, D. (2010). Digitalno učenje i obrazovne tehnologije. Novi Sad: Filozofski fakultet.

[6] Clark, R. C., & Mayer, R. E. (2016). *E-Learning and the Science of Instruction* (4th ed.). Hoboken: Wiley.

Received: November 4, 2025
Accepted: November 13, 2025

ABOUT THE AUTHORS

Karlo Čuković-Tkalčec (1991, Koprivnica) graduated from Fran Galović Gymnasium, where he actively participated in field-teaching projects and extracurricular activities, independently producing video recordings and film editing for school presentations. After completing vocational retraining, he obtained an instructor's license and worked for five years as a driving instructor. In 2020, he enrolled at the Pan-European University Apeiron, where he completed the first cycle of teacher-training studies in informatics in 2023 (180 ECTS), followed by the first cycle of academic studies in computer science and informatics in 2024 (240 ECTS), graduating with excellent results; he is currently completing a second-cycle master's program (300 ECTS). From an early age, he has been developing strong computer-related skills, which he further enhanced through his studies and professional practice at the gymnasium by creating teaching materials, presentations, quizzes, and exams, and by working directly with students. In addition to his work in a driving school, he has held various occasional jobs in production facilities and forest maintenance. Since September 2025, he has been employed as a mathematics teacher at Koprivnički Bregi Elementary School and at Molve Elementary School.

FOR CITATION

Karlo Čuković - Tkalčec, Model to improve distance learning system LOOMEN, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 15(2025)2:161-172, (UDC: 004.738.5:37.018.43), (DOI: 10.7251/JIT2502161T), Volume 15, Number 2, Banja Luka, December (81-176), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004

JITA

Journal of Information Technology and Applications
Pan-European University APEIRON Banja Luka
https://www.jita-au.com
jita@apeiron-edu.eu

PUBLISHER: **Pan-European University APEIRON**, Banja Luka
College of Information Technology Banja Luka, Republic of Srpska, BiH
www.apeiron-uni.eu
Darko Uremović, Person Responsible for the Publisher
Aleksandra Vidović, PhD, Editor of University Publications

ISSN 2232-9625