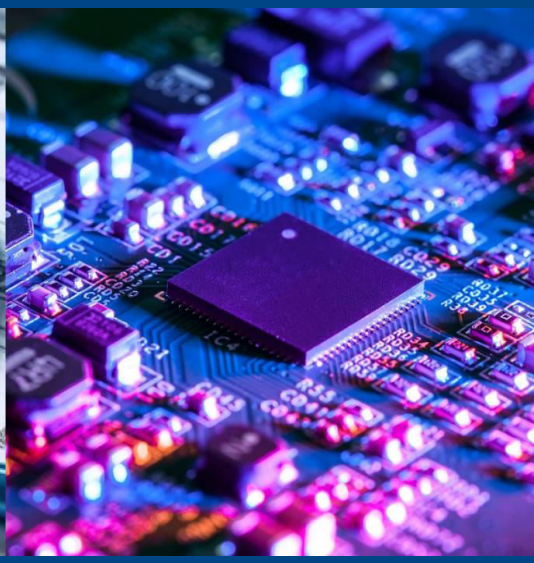
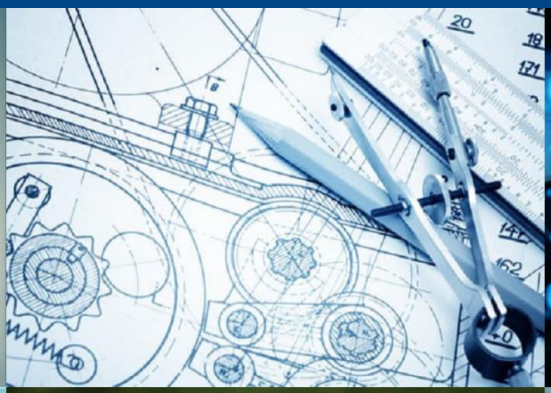
ISSN: 2582-7219

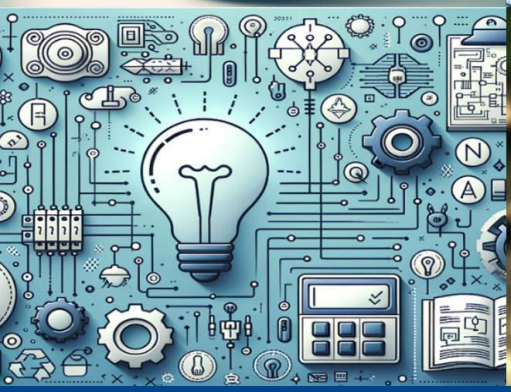




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.864

Volume 9, Issue 5, May 2026

© 2026 IJMRSET | Volume 9, Issue 5, May 2026| DOI:10.15680/IJMRSET.2026.0905027

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Edge Artificial Intelligence (Edge AI)

Darshan M S[1], Mr. Naseerhusen Ankalagi [2]

PG Student, Dept. of MCA, City Engineering College, Bengaluru, Karnataka, India[1]

Assistant Professor, Dept. of MCA, City Engineering College, Bengaluru, Karnataka, India[2]

ABSTRACT: Edge Artificial Intelligence represents a significant evolution in the field of artificial intelligence by enabling data processing and intelligent decision-making directly at or near the data source instead of depending on cloud infrastructures that are centralized. The proliferation of Internet of Things devices and real-time applications, traditional cloud-based AI systems face critical challenges such as high latency, bandwidth limitations, increased operational costs, and privacy concerns. Edge AI addresses these limitations by shifting computation to edge devices such as sensors, smartphones, embedded systems, and edge gateways, thereby reducing the dependency on cloud communication and enabling faster response times.

This paper presents a comprehensive analysis of Edge AI, focusing on its system architecture, communication efficiency, security considerations, and performance outcomes. The proposed system model consists of a multi-layered architecture that includes edge devices for data acquisition, edge nodes for local processing, and cloud servers for model training and global updates. The study assumes constraints such as limited computational resources, energy efficiency requirements, and intermittent network connectivity, which are common in real-world edge environments. To overcome these challenges, various optimization to guarantee effective deployment of machine learning models on devices with limited resources, strategies like model compression, quantization, and pruning are used.

I. INTRODUCTION

The rapid advancement of digital technologies and the exponential growth of data generated by connected devices have profoundly changed the artificial intelligence landscape and computing systems. In recent years, the proliferation of Internet of Things devices, including smart sensors, wearable devices, surveillance cameras, and autonomous systems, has led to an unprecedented increase in data generation at the network edge. Traditional cloud-based AI architectures, which rely on centralized data processing and storage, are increasingly facing challenges in handling this massive volume of data efficiently high latency, higher bandwidth usage, data privacy issues, and reliance on steady network connectivity are some of these difficulties. A decentralized technique that can manage information closer to its source while preserving efficiency and dependability is therefore becoming more and more necessary.Edge Artificial Intelligence has become a viable way to overcome these restrictions.

Edge AI refers to the deployment of artificial intelligence algorithms directly on edge devices, enabling real-time data processing and intelligent decision-making without the need to transmit data to distant cloud servers. By bringing Edge AI ensures quicker response times and lowers latency by computing closer to the data source, which are essential for time-sensitive applications such as autonomous driving, healthcare monitoring, industrial automation, and smart surveillance systems. This shift from centralized to distributed intelligence represents a fundamental transformation in how AI systems are designed and deployed.

The requirement for real-time processing skills is one of the main drivers of Edge AI. Data processing delays can have serious repercussions in applications like crowd surveillance, anomaly detection, and emergency response systems. For instance, in a crowd monitoring system like the proposed “Crowd Guard AI,” detecting panic behaviour or abnormal movement patterns requires immediate analysis and response. Relying solely on cloud-based processing introduces latency due to data transmission and network delays, which can hinder timely intervention. Edge AI addresses this issue by enabling on-device inference, allowing systems to process video streams and sensor data locally and generate instant alerts.

© 2026 IJMRSET | Volume 9, Issue 5, May 2026| DOI:10.15680/IJMRSET.2026.0905027

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. SYSTEM MODEL AND ASSUMPTIONS

The Edge AI system is designed as a distributed architecture consisting of three main layersThe cloud, edge devices, and edge nodes. Sensors, cameras, and cell phones are examples of edge devices that are in charge of gathering real-time environmental data. These devices may perform basic preprocessing tasks before sending the data to edge nodes. The edge nodes or gateways act as intermediate processing units with higher computational power, where real-time inference and prompt decision-making are carried out using machine learning models. The cloud layer is mainly used for large-scale data storage, model training, and system management. Trained models are periodically deployed back to the edge, ensuring a balance between high computational capability and low-latency processing.
The system operates under several key assumptions. Edge devices are resource-constrained with constrained energy, memory, and processing capacity, necessitating lightweight and optimized AI models. The system should continue to operate effectively even in the event of erratic or restricted network connectivity minimal cloud interaction. Data generated is distributed and heterogeneous, and privacy is a major concern; hence, sensitive data is processed locally whenever possible using techniques such as federated learning. Additionally, the system must support scalability and real-time responsiveness to handle increasing data loads and ensure timely decision-making in critical applications.

III. EFFICIENT COMMUNICATION

Efficient communication is a critical component of Edge AI systems, as continuous transmission of large volumes Sending raw data to the cloud could lead to significant bandwidth consumption, network congestion, and increased delay. Through local data processing and inference at edge nodes or devices, Edge AI reduces data transfer. Methods like data compression, feature extraction, quantization, and model pruning are used to reduce the size of data before transmission. Instead of sending entire datasets, only relevant insights or summarized information are communicated to the cloud, greatly increasing network effectiveness and cutting expenses.

Another important approach to efficient communication is the use of collaborative and distributed learning techniques such as federated learning. In this method, multiple edge devices train a shared model locally and only transmit model updates to a central server, rather than sharing raw data. This improves data privacy while also lowering transmission overhead.Additionally, edge-cloud collaboration strategies are implemented, where time-sensitive tasks are handled at the edge while computationally intensive processes are offloaded to the cloud. This balanced distribution of workload ensures optimal utilization of resources while maintaining low latency and high system performance.

IV. SECURITY

Security is a fundamental concern in Edge Artificial Intelligence (Edge AI) systems due to their distributed and decentralized nature. Unlike traditional cloud-based systems, where data is processed in a centralized and controlled environment, Edge AI involves multiple devices operating at different locations, often in untrusted or physically exposed environments. This increases the risk of several security risks, such as adversarial attacks, model modification, illegal access, and data breaches. Ensuring the security, integrity, and availability of sensitive data becomes crucial since Internet of Things nodes, sensors, and cameras are examples of edge devices that are constantly gather and analyse information. The primary security challenges in Edge AI is data privacy. In many applications, such as healthcare monitoring, surveillance, and smart cities, the data being processed may contain sensitive personal or organizational information. Transmitting such data to the cloud can expose it to interception or misuse. Edge AI addresses this issue by processing data locally on devices, thereby reducing the need to transfer raw data over networks. Additionally, techniques such as data encryption are used to safeguard data while it is being transmitted and stored. Encryption ensures that even if data is intercepted, it cannot be easily accessed or understood by unauthorized entities.
Another major threat in Edge AI systems is adversarial attacks, in which malevolent parties alter input data in order to trick machine learning models. For instance, slight modifications to an image can cause a model to misclassify objects, leading to incorrect decisions. Similarly, data poisoning attacks can occur In order to tamper with the model's learning process, attackers introduce fake data during the training phase. To mitigate these risks, robust model validation techniques, anomaly detection mechanisms, and secure training protocols are implemented. Regular updates and monitoring of models also help in identifying and preventing such attacks.

V. RESULT AND DISCUSSION

The implementation and evaluation of Edge Artificial Intelligence (Edge AI) systems demonstrate significant improvements in performance, efficiency, and reliability when compared to traditional cloud-based approaches. One of the most notable outcomes is the reduction in latency, as data processing and inference are performed directly at the edge devices or nearby edge nodes. This enables real-time or near real-time decision-making, which is essential for applications such as crowd monitoring, autonomous vehicles, smart surveillance, and healthcare systems. Experimental observations indicate that response times are significantly faster since the dependency on network communication with distant cloud servers is minimized.

Another important result is the reduction in bandwidth usage. In conventional cloud-based systems, large volumes of raw data must be continuously transmitted to centralized servers for processing, leading to network congestion and increased operational costs. Edge AI addresses this issue by processing data locally and transmitting only relevant information or summarized results to the cloud. Techniques such as data filtering, compression, and feature extraction further optimize communication efficiency. As a result, the overall network load is reduced, making Edge AI systems more suitable for environments with limited or unstable connectivity.

Additionally, edge AI systems exhibit increased robustness and dependability, especially in situations where network access is sporadic or non-existent. The system keeps working even when there is a network outage since edge devices may function independently without regular cloud involvement. Because of this, Edge AI is very useful in remote or crucial settings like industrial automation, disaster management, and defined applications. continued operation improves system reliability and guarantees continued service.

VI. CONCLUSION

Edge Artificial Intelligence represents a major advancement in modern computing by enabling data process and intelligent decision-making is closer to the data source. By reducing reliance on centralized cloud systems, Edge AI significantly lowers latency, minimizes bandwidth usage, and enhances data privacy. These advantages make it highly suitable for real-time applications such as smart surveillance, healthcare monitoring, autonomous systems, and industrial automation. The integration of edge computing with AI also improves system reliability and scalability, allowing efficient handling of large volumes of distributed data.

Despite its benefits, Edge AI faces challenges such as limited computational resources, energy constraints, and security risks in decentralized environments. However, with advancements in lightweight AI models, edge hardware, and technologies like federated learning and 5G, these challenges are gradually being addressed. Overall, Edge AI is a transformative technology that overcomes the limitations of traditional cloud-based systems and is expected to play a key role in the future of intelligent, real-time, and secure applications.

REFERENCES

1. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, Oct. 2016.
2. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, May 2015.
3. K. Zhang, Y. Mao, S. Leng, A. Vinel, and Y. Zhang, “Delay constrained offloading for mobile edge computing in cloud-enabled vehicular networks,” IEEE ICC, pp. 1–6, 2016.
4. X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, “Convergence of edge computing and deep learning: A comprehensive survey,” IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 869–904, 2020.
5. S. Li, L. D. Xu, and S. Zhao, “The Internet of Things: A survey,” Information Systems Frontiers, vol. 17, no. 2, pp. 243–259, 2015.
6. T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, “On multi-access edge computing: A survey of the emerging 5G network edge architecture,” IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1657–1681, 2017.
7. J. Ren, G. Yu, Y. Cai, and Y. He, “Latency optimization for resource allocation in mobile-edge computation offloading,” IEEE Transactions on Wireless Communications, vol. 17, no. 8, pp. 5506–5519, Aug. 2018.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

8. H. Brendan McMahan et al., "Communication-efficient learning of deep networks from decentralized data," Proc. AISTATS, pp. 1273–1282, 2021.
9. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2023.
10. N. D. Lane et al., "DeepX: A software accelerator for low-power deep learning inference on mobile devices," IPSN, pp. 1–12, 2024.
11. M. Satyanarayanan, "The emergence of edge computing," Computer, vol. 50, no. 1, pp. 30–39, Jan. 2024.
12. Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322–2358, 2025.
13. P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1628–1656, 2025.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com