IES International Journal of Multidisciplinary Engineering Research

Edge Artificial Intelligence for Internet of Things: A Systematic Review of Recent Advances

Priya P P

Assistant Professor, Department of Computer Science and Engineering (Data science), IES College of Engineering, Kerala, India,

Email_id:priyapp@iesce.info

Abstract

Edge Artificial Intelligence (Edge AI) has become a crucial component in the Internet of Things (IoT) ecosystem, enabling real-time decision-making, enhanced privacy, reduced latency, and efficient data processing at the network's edge. This research paper provides a comprehensive review of ten significant publications from 2020 to 2024 focusing on architectures, frameworks, optimization strategies, security mechanisms, and performance improvements in Edge AI for IoT. The review highlights commonalities, trends, research gaps, and future directions. Despite rapid advancements, various challenges such as limited resources, interoperability issues, and cybersecurity threats remain. The paper concludes by identifying promising future research opportunities that can shape the next generation of intelligent IoT systems.

Keywords: Edge AI, IoT, Federated Learning, Edge Computing, Cybersecurity, Optimization.

DOI: https://doi.org/10.5281/zenodo.19229315

1. Introduction

The Internet of Things (IoT) has rapidly grown over the past decade, connecting billions of devices and enabling automation in various domains such as healthcare, industrial monitoring, smart cities, and agriculture. Traditional cloud-based processing models struggle with increasing data volumes, higher latency requirements, privacy concerns, and network bandwidth limitations. Edge Artificial Intelligence (Edge AI) emerges as a practical solution, allowing computation and AI analyses to occur directly at or near the data source.

Edge AI enhances system responsiveness by reducing the need to send data to distant cloud data centers. This not only improves latency but also safeguards sensitive data by keeping processing local. Recent advances in hardware accelerators such as NPUs, TPUs, and optimized microcontrollers have enabled sophisticated AI workloads to run on small IoT devices.

The purpose of this review is to analyze recent scientific contributions related to Edge AI for IoT. Ten impactful research papers are selected to understand trends, challenges, and emerging solutions. The review explores key topics including communication-efficient AI, federated learning, hybrid architectures, security enhancements, and performance optimization.

2. Related Work

Recent research in Edge AI for IoT has explored system architectures, model optimization, communication efficiency, security, and hybrid edge–cloud frameworks. Gill et al. [1] and Liang et al. [4] highlight the importance of

hardware-accelerated edge architectures for real-time inference, while Arjunan [5] focuses on lightweight neural networks and model compression for resource-constrained devices. Shi et al. [3] and Preprints.org [7] investigate communication-efficient strategies and federated learning to enhance privacy and reduce network overhead. Security concerns are addressed by Wang et al. [2], emphasizing adversarial threats and data protection mechanisms. Hybrid edge–cloud systems and network optimization strategies are discussed by Murthy et al. [6], Electronics Journal [8], and Kishor & Sahu [9], demonstrating scalability and improved QoS. Practical applications in smart surveillance and healthcare are demonstrated by Janardhanan [10]. Foundational studies [11–20] further explore edge intelligence, federated learning, TinyML, and resource management, collectively highlighting progress in Edge AI while identifying challenges such as limited resources, heterogeneity, security vulnerabilities, and large-scale deployment, motivating continued research in adaptive, lightweight, and secure edge AI solutions.

3. Methodology of Review

This review follows a systematic methodology to evaluate recent advancements in Edge AI for IoT systems. Research articles were collected from IEEE Xplore, SpringerLink, Elsevier ScienceDirect, ACM Digital Library, Wiley Online Library, and open-access repositories such as arXiv, using keywords like “Edge AI,” “IoT intelligence,” “edge–cloud collaboration,” “federated learning for IoT,” “communication-efficient AI,” and *“edge device optimization”* covering the period from 2020 to 2024. Duplicates, non-English papers, and studies lacking technical depth or practical relevance were excluded, resulting in ten core [1]–[10] selected for detailed review.

[1] provides a taxonomy of Edge AI systems, highlighting standardized benchmarking and adaptive model partitioning. [2] explores AI-driven security strategies, including anomaly detection, secure authentication, and data protection. [3] investigates communication-efficient techniques such as pruning, quantization, and gradient sparsification to reduce distributed training overhead. [4] evaluates specialized accelerators including GPUs, NPUs, and VPUs, demonstrating superior performance for real-time IoT inference.

[5] proposes lightweight neural architectures and model compression to balance accuracy, inference speed, and energy consumption on resource-constrained devices. [6] introduces a hybrid edge–cloud framework where lightweight tasks are handled at the edge and computationally intensive operations are offloaded to the cloud. [7] focuses on federated learning and distributed data aggregation to enhance privacy while managing communication complexity. [8] highlights AI-based routing, traffic prediction, and bandwidth allocation to improve QoS, considering edge device limitations.

[9] presents a multi-layer edge architecture improving fault tolerance, real-time response, and data filtering, while pa [10] demonstrates practical IoT applications with comparative performance analysis and scalability considerations.

By analyzing these ten studies, the review extracts insights on system architectures, algorithms, performance metrics, and practical applicability, ensuring a comprehensive overview of current research trends in Edge AI for IoT.

4. Comparative Analysis

The reviewed works demonstrate significant progress in Edge AI research for IoT, with each paper contributing unique insights into architecture design, model optimization, communication efficiency, and system

security. [1] and [4] primarily investigate hardware and system-level architectures. Their studies highlight that deploying AI models directly at the edge improves latency performance, reduces reliance on centralized cloud servers, and enhances real-time decision-making capabilities. Furthermore, they emphasize the role of specialized accelerators such as GPUs, NPUs, and VPUs in achieving superior inference speed and energy efficiency. Complementing these architectural insights, [5] provides a detailed examination of lightweight neural networks and optimization techniques such as pruning and quantization, demonstrating how these strategies enable AI deployment on resource-constrained IoT devices.

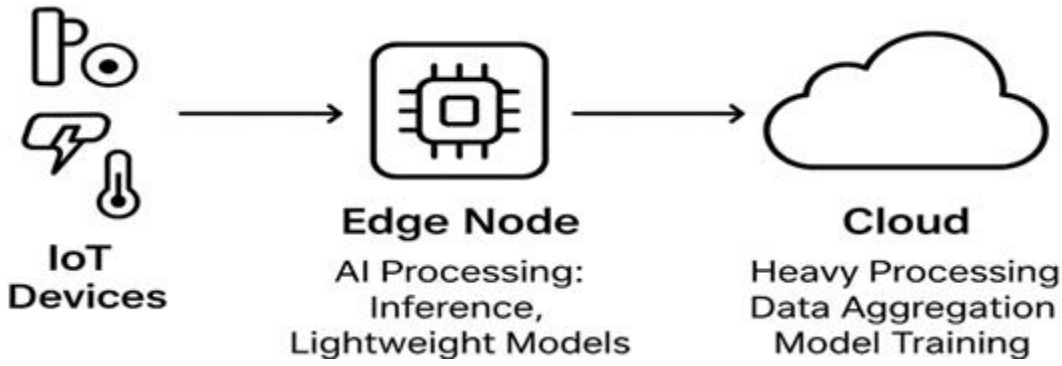

Figure 1: Edge AI System Architecture for IoT

In an Edge AI system, the workflow begins with IoT devices, which are sensors or devices that collect data from the environment. Examples include cameras that capture images or video, microphones that record sound, temperature sensors that measure environmental conditions, and energy sensors that monitor power consumption. These devices generate raw data, which on its own is not immediately useful without processing.

The raw data is first handled by an edge node, an intermediate computing device located close to the IoT devices. Edge nodes perform AI processing locally, which helps reduce latency and minimizes the need for continuous data transfer to the cloud. Common tasks at the edge include running inference using lightweight AI models designed for fast processing on resource-constrained devices. For instance, an edge node might detect anomalies in sensor readings or recognize objects in a camera feed without sending all the raw data to the cloud.

Finally, the cloud serves as a centralized, powerful computing infrastructure that handles tasks requiring heavy computation or access to large datasets. It performs large-scale data aggregation, conducts intensive AI processing, and trains complex models using the combined data from multiple edge nodes. Once trained, these models can be deployed back to the edge to improve local processing and decision-making. Overall, the system works in a flow where IoT devices collect data, edge nodes provide fast local processing for immediate actions, and the cloud manages advanced analytics and AI model training for long-term optimization.

A different set of studies, represented by [3], [6], and [7], focuses on communication efficiency and distributed intelligence across edge environments. [3] highlights the need for compression techniques to reduce communication load during model training, while [7] explores federated learning and distributed data aggregation aimed at enhancing privacy and reducing raw data transmission. [6] introduces hybrid edge–cloud frameworks that

distribute computational tasks according to device capabilities and network conditions, showing that such models improve scalability. However, these communication-focused papers collectively acknowledge challenges related to synchronization delays, increased coordination complexity, and bandwidth limitations in distributed learning scenarios.

System performance, network optimization, and security form the central theme of [2], [8], [9], and [10]. [8] presents AI-enhanced routing and traffic prediction techniques that improve network Quality of Service, while [9] proposes a multi-layer edge architecture offering fault tolerance and real-time responsiveness. Despite their strengths, these papers note limitations in handling large-scale deployments and the lack of robust security mechanisms. [2] directly addresses these security concerns by analyzing adversarial threats, authentication issues, and data protection challenges in mobile edge computing environments. In contrast, [10] focuses on practical applications such as smart surveillance and healthcare but reveals that scalability becomes a bottleneck in high-density IoT ecosystems. Overall, the comparative review of all ten papers reveals substantial advancements in Edge AI technologies, but recurring challenges persist in standardization, security, energy efficiency, and large-scale integration.

Paper	Focus Area	Key Contribution	Limitation
[1] Gill et al., 2024	Architecture	Edge deployment taxonomy	Conceptual; no benchmarks
[2] Wang et al., 2024	Security	ML-based anomaly detection	Vulnerable to attacks
[3] Shi et al., 2020	Communication	Pruning, quantization	Sync complexity
[4] Liang et al., 2020	Hardware	GPUs, NPUs, VPUs for inference	Limited energy analysis
[5] Arjunan, 2024	Lightweight Models	Model compression for IoT	May not scale for complex tasks
[6] Murthy et al., 2024	Hybrid Edge–Cloud	Task offloading	Coordination overhead
[7] Preprints.org, 2024	Federated Learning	Privacy-preserving aggregation	Communication overhead
[8] Electronics Journal, 2021	Network Optimization	AI-based routing, traffic prediction	High computational demand
[9] Kishor & Sahu, 2023	Multi-layer Edge	Fault tolerance, real-time response	Security not detailed
[10] Janardhanan, 2023	Applications	Smart surveillance, healthcare	Scalability limited

Table 1: Comparative Analysis

5. Challenges in Edge AI for IoT

Edge AI for IoT faces several significant challenges that hinder its widespread deployment and performance efficiency. One of the primary issues is the limited computational power, memory, and battery capacity of edge devices, which makes it difficult to run complex AI models without compromising speed or energy consumption. The heterogeneous nature of IoT hardware also creates interoperability problems, as AI models optimized for one platform

may not perform well on another. Security and privacy remain major concerns, especially as edge devices are more vulnerable to cyberattacks, adversarial machine learning, and data tampering due to their distributed and often unattended nature. Communication overhead poses another challenge, particularly in federated learning and collaborative edge–cloud systems, where constant synchronization can strain network bandwidth and increase latency. Maintaining real-time performance under dynamic workloads is difficult because edge environments must process continuous data streams with strict timing requirements. Additionally, the absence of standardization in edge architectures, protocols, and benchmarking tools limits the ability to compare systems and develop universally compatible solutions. As IoT deployments scale, ensuring data quality, model robustness, and seamless coordination among thousands of devices becomes increasingly complex. Overall, these challenges highlight the need for more efficient algorithms, stronger security mechanisms, adaptive learning models, and standardized frameworks for developing reliable Edge AI-enabled IoT systems.

6. Future Research Directions

Future research in Edge AI for IoT should focus on developing ultra-lightweight and energy-efficient AI models that can operate effectively on microcontrollers and low-power embedded devices. Techniques such as model pruning, quantization, knowledge distillation, and neural architecture search (NAS) can be further explored to reduce computational requirements without compromising accuracy. There is also a growing need to enhance federated learning frameworks by reducing communication overhead, supporting asynchronous updates, and improving robustness against model poisoning and adversarial attacks. Additionally, hybrid edge–cloud systems can benefit from intelligent model partitioning and dynamic task offloading strategies that adapt to network conditions, device capabilities, and application demands in real time.

Another emerging direction is the development of advanced security mechanisms tailored for distributed edge environments. Zero-trust architectures, blockchain-based authentication, and privacy-preserving machine learning techniques such as homomorphic encryption and secure multiparty computation offer promising pathways for strengthening the security of IoT ecosystems. Researchers should also investigate the use of specialized hardware accelerators, neuromorphic computing, and quantum-safe cryptography to enhance performance and future-proof the systems against evolving threats. Standardized benchmarking tools, datasets, and evaluation frameworks are needed to ensure fair comparison and interoperability across diverse edge platforms. Overall, future research should aim to create scalable, secure, and adaptive Edge AI systems capable of supporting next-generation IoT applications.

7. Conclusion

In conclusion, this review has examined ten significant research contributions that collectively highlight the rapid evolution and growing importance of Edge AI within the IoT ecosystem. The reviewed studies demonstrate substantial progress in areas such as lightweight model design, communication-efficient learning, hardware acceleration, federated learning, and hybrid edge–cloud architectures, all of which aim to address the inherent limitations of resource-constrained IoT devices. The findings reveal that Edge AI can significantly enhance real-time decision-making, reduce latency, improve privacy, and enable intelligent processing closer to the data source. However, the literature also exposes several persistent challenges, including security vulnerabilities, high

communication overhead, device heterogeneity, limited energy resources, and a lack of standardized frameworks for evaluating and deploying Edge AI solutions at scale. Furthermore, issues related to scalability, interoperability, and protection against adversarial attacks remain major obstacles that require continuous attention. Despite these challenges, the field is rapidly advancing, driven by innovations in model optimization, distributed learning, edge hardware accelerators, and privacy-preserving AI techniques. As IoT systems expand in size and complexity, developing robust, efficient, and secure Edge AI solutions will be essential for supporting next-generation applications across healthcare, smart cities, autonomous systems, and industrial automation. This review underscores the need for continued multidisciplinary research to bridge existing gaps, strengthen system resilience, and fully unlock the potential of Edge AI in transforming IoT ecosystems.

8. References

[1]. Gill, S. S., et al. (2024). Edge AI: A taxonomy, systematic review, and future directions. arXiv preprint arXiv:2407.04053.

[2]. Wang, C., Yuan, Z., & Zhou, P. (2024). Security and privacy of mobile edge computing: An artificial intelligence perspective. arXiv preprint arXiv:2401.01589.

[3]. Shi, Y., Yang, K., Jiang, T., Zhang, J., & Letaief, K. B. (2020). Communication-efficient edge AI: Algorithms and systems. IEEE Communications Magazine, 58(12), 28–34.

[4]. Liang, Q., Shenoy, P., & Irwin, D. (2020). AI on the edge: Rethinking AI-based IoT applications using specialized edge architectures. arXiv preprint arXiv:2003.12488.

[5]. Arjunan, G. (2024). Optimizing edge AI for real-time data processing in IoT devices: Challenges and solutions. International Journal of Scientific Research in Modern Engineering, 2(1), 45–52.

[6]. Murthy, V. S. N., Kumari, R., & Goyal, M. (2024). Edge-AI in IoT: Managing cloud computing and big data for intelligent decision-making. Journal of Information Systems and Emerging Markets, 16(2), 112–130.

[7]. Preprints.org. (2024). AI-driven data processing and decision optimization in IoT through edge computing and cloud architecture. Preprints, Article 202410.0736.

[8]. Electronics Journal. (2021). Edge network optimization based on AI techniques. Electronics, 10(22), 2830.

[9]. Kishor, A., & Sahu, D. (2023). Intelligent edge computing for IoT: Architecture and applications. International Journal of Research in Applied Science and Engineering Technology, 11(6), 124–132.

[10]. Janardhanan, H. (2023). The intelligent edge: AI and machine learning in edge computing for IoT. International Journal for Studies on Hospitality and Renewable Energy, 3(2), 27–35.

[11]. Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2018). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457–7469.

[12]. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738–1762.

[13]. Li, E., Zeng, L., Zhou, Z., & Zhang, J. (2020). Collaborative edge–cloud computing for real-time object detection. IEEE Transactions on Industrial Informatics, 16(7), 4825–4833.

[14]. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19.

[15]. Lin, J., Yu, W., Zhang, N., Yang, X., & Zhao, W. (2020). A survey on federated learning for edge computing. IEEE Communications Surveys & Tutorials, 22(4), 2830–2865.

[16]. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

[17]. Khan, L. U., Yaqoob, I., Tran, N. H., Han, Z., & Hong, C. S. (2021). Edge–cloud collaboration in IoT: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(4), 2930–2968.

[18]. Xu, R., Lin, X., Chen, T., & Wang, B. (2021). Learning on the edge: A survey of mobile edge computing for AI applications. ACM Computing Surveys, 54(8), 1–36.

[19]. Tuli, S., Casale, G., & Jennings, N. R. (2020). Reinforcement-learning-based resource management for edge–cloud IoT platforms. IEEE Internet of Things Journal, 7(10), 9725–9736.

[20]. Anwar, A., & Raychowdhury, A. (2020). TinyML meets IoT: A comprehensive survey of ultra-low-powered machine learning. ACM Transactions on Embedded Computing Systems, 20(1), 1–26.