International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. 11 Special Issue 02 | 2026

A Review – Comparative study of Cloud and Edge AI

Mr. Sudesh L. Farpat[1]
[1] Research Scholar, Department, Computer Science and Engineering, SKU University, MP, India

DOI: 10.5281/zenodo.19254086

ABSTRACT

Artificial Intelligence (AI) has emerged as a foundational technology for modern data-driven systems, enabling intelligent decision-making across diverse domains such as healthcare, industrial automation, smart cities, and autonomous systems. Traditionally, AI workloads have been executed in centralized cloud infrastructures, which offer vast computational power, scalable storage, and advanced machine learning platforms capable of training and deploying complex deep learning models. Cloud-based AI has been widely adopted due to its flexibility, cost-effectiveness through pay-as-you-go models, and ease of managing large-scale data analytics. However, the increasing proliferation of Internet of Things (IoT) devices has led to an unprecedented surge in real-time data generation at the network edge, exposing significant limitations of cloud-centric AI architectures.

Latency-sensitive applications, such as remote patient monitoring, real-time fault detection in industrial systems, and autonomous control, require rapid decision-making that cannot always tolerate the communication delays associated with transmitting data to distant cloud data centers. Additionally, concerns related to bandwidth consumption, intermittent network connectivity, data privacy, and regulatory compliance further challenge the feasibility of relying solely on cloud-based AI. These limitations have driven the emergence of Edge AI, a paradigm in which AI inference and data processing are performed closer to the data source, such as on embedded devices, edge gateways, or local servers.

This review paper presents a comprehensive analysis of Edge AI and Cloud AI architectures, examining their fundamental design principles, supporting frameworks, operational advantages, and inherent limitations. The paper further explores their applicability across key domains, including healthcare and industrial IoT, highlighting how each paradigm addresses performance, scalability, and security requirements. Moreover, this review discusses emerging hybrid edge–cloud models, open research challenges, and future directions, providing valuable insights to guide researchers and practitioners in selecting, designing, and integrating effective AI deployment strategies for next-generation intelligent systems.

Keywords: - Edge AI, Cloud AI, Artificial Intelligence, Internet of Things, Low Latency, Distributed AI

1. INTRODUCTION

Artificial Intelligence (AI) has significantly transformed a wide range of sectors, including healthcare, manufacturing, transportation, and smart cities, by enabling intelligent automation, predictive analytics, and data-driven decision-making. Among various deployment paradigms, Cloud AI has long been the dominant approach, offering access to powerful computational resources, centralized data management, and scalable platforms capable of handling large volumes of data. Cloud-based AI systems support the training and deployment of complex deep learning models and provide flexibility through elastic resource allocation. As a result, they have been widely adopted for applications requiring extensive data processing and long-term analytics.

However, the rapid proliferation of Internet of Things (IoT) devices has led to the generation of massive real-time data streams at the network edge. This shift has exposed several limitations of cloud-centric AI systems, particularly for latency-sensitive applications. Transmitting raw data to distant cloud data centers introduces communication delays, increased bandwidth consumption, and dependence on reliable network connectivity. Furthermore, concerns related to data privacy, security, and regulatory compliance have become increasingly prominent, especially in domains such as healthcare and industrial automation where sensitive information is involved.

To address these challenges, Edge AI has emerged as a promising paradigm that enables AI inference and data processing to be performed at or near the data source, such as on sensors, embedded devices, or edge gateways. By minimizing data transmission and enabling localized decision-making, Edge AI significantly reduces latency, improves system responsiveness, and enhances data privacy. This review paper surveys both Edge AI and Cloud AI paradigms, providing a comparative analysis of their architectures, performance characteristics, advantages, and limitations. The paper aims to offer a structured understanding of the complementary roles of Edge AI and Cloud AI within modern AI ecosystems and to guide researchers and practitioners in selecting appropriate AI deployment strategies for emerging intelligent applications.

2. CLOUD AI ARCHITECTURE - EDGE VS CLOUD

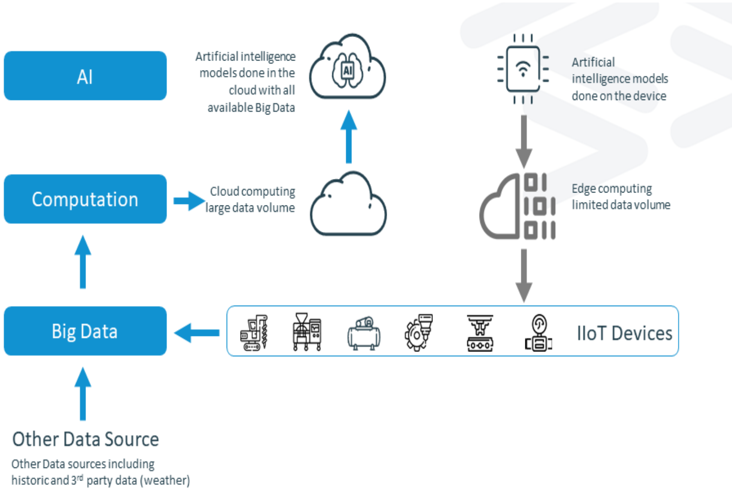

Figure – Edge vs Cloud – Stored and Analyzed

Cloud AI refers to AI systems deployed on centralized cloud infrastructure where data is transmitted from end devices to remote servers for processing. The diagram illustrates the functional and architectural differences between Cloud AI and Edge AI in an Industrial Internet of Things (IIoT) environment.On the left side, the diagram represents the Cloud AI workflow. Data generated by IIoT devices is transmitted to centralized cloud infrastructure, where it is combined with big data from other sources such as historical records and third-party data (e.g., weather or external databases). This large data volume is processed using cloud computing resources that provide high computational power and scalability. AI models are trained and executed in the cloud, leveraging extensive datasets to achieve high accuracy. While this approach supports complex analytics and long-term insights, it introduces challenges such as higher latency, increased bandwidth usage, and dependency on network connectivity.

On the right side, the diagram depicts the Edge AI approach. Here, AI models are deployed directly on edge devices or edge computing nodes close to IIoT devices. Data is processed locally with a limited data volume, enabling faster inference and real-time decision-making. This reduces latency, minimizes data transmission to the cloud, and enhances privacy and reliability, particularly in mission-critical industrial applications.

Overall, the diagram highlights the complementary nature of Cloud AI and Edge AI: cloud-based systems excel in large-scale data processing and model training, while edge-based systems are optimized for low-latency, real-time intelligence at the device level.

2.1 Characteristics of Cloud AI

- High computational and storage capacity
- Support for complex deep learning models
- Centralized training and deployment
- Elastic scalability

2.2 Advantages

- Efficient handling of large datasets
- Advanced model training capabilities
- Simplified maintenance and updates

2.3 Limitations

- Network-dependent latency
- Privacy and data security risks
- High bandwidth consumption

3. EDGE AI ARCHITECTURE

Edge AI moves AI inference closer to data sources, such as sensors, embedded devices, or edge gateways.

3.1 Characteristics of Edge AI

- Localized processing

- Reduced data transmission
- Real-time inference

3.2 Advantages

- Ultra-low latency
- Improved privacy
- Reduced bandwidth usage
- Increased system reliability

3.3 Limitations

- Limited hardware resources
- Model optimization requirements
- Deployment and management complexity

4. EDGE AI AND CLOUD AI FRAMEWORKS

4.1 Edge AI Frameworks

- TensorFlow Lite
- ONNX Runtime
- OpenVINO
- NVIDIA TensorRT

These frameworks enable efficient inference on resource-constrained devices.

4.2 Cloud AI Platforms

- AWS SageMaker
- Google Cloud AI Platform
- Microsoft Azure Machine Learning

Cloud platforms support end-to-end AI lifecycle management.

5. COMPARATIVE ANALYSIS OF EDGE AI AND CLOUD AI

Feature	Edge AI	Cloud AI
Latency	Very low	Network-dependent
Scalability	Limited	High
Privacy	High	Moderate
Compute Power	Limited	High
Energy Efficiency	Optimized locally	Centralized
Deployment Cost	Hardware-dependent	Usage-based

6. APPLICATION DOMAINS

6.1 Healthcare

- Remote patient monitoring
- Medical imaging
- Wearable health devices

6.2 Industrial IoT

- Predictive maintenance
- Fault detection
- Smart manufacturing

6.3 Smart Cities

- Traffic monitoring
- Surveillance systems
- Energy management

7. HYBRID EDGE–CLOUD AI

Hybrid architectures combine Edge AI for real-time inference and Cloud AI for training, coordination, and large-scale analytics.

Benefits:

- Balanced latency and scalability
- Improved system intelligence
- Reduced cloud workload

8. CONCLUSION

This review paper highlights the complementary nature of Edge AI and Cloud AI within modern artificial intelligence ecosystems. Cloud AI continues to play a vital role in large-scale data processing, complex model training, and centralized analytics due to its vast computational resources and scalability. In contrast, Edge AI is essential for real-time, latency-sensitive, and privacy-critical applications, as it enables localized data processing and rapid decision-making at the data source. By reducing reliance on continuous cloud connectivity, Edge AI enhances system responsiveness and reliability. Hybrid edge–cloud approaches, which combine the strengths of both paradigms, represent the most promising direction for future AI systems, enabling intelligent, scalable, and efficient deployment strategies across diverse application domains.

9.REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[2] M. Satyanarayanan, "The Emergence of Edge Computing," *IEEE Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[3] S. Shi, V. Gupta, E. H. M. Sha, and Z. Wang, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[4] W. Shi and S. Dustdar, "The Promise of Edge Computing," *IEEE Computer*, vol. 49, no. 5, pp. 78–81, 2016.

[5] J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[6] K. Zhang, Y. Mao, S. Leng, Y. He, and Y. Zhang, "Energy-Efficient AI at the Edge: A Survey," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 26–31, 2019.

[7] X. Xu, W. Yu, D. Griffith, N. Golmie, and P. Moulema, "Toward Integrating Edge Computing and Cloud Computing," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 8–15, 2019.

[8] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2961–2991, 2018.

[9] H. Qi, Z. Chen, and S. Yang, "Deep Learning Applications in Industrial Internet of Things: A Survey," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 102–115, 2020.

[10] A. Abbas, S. U. Khan, and M. Alam, "Edge Computing for Healthcare IoT: A Survey," *IEEE Access*, vol. 8, pp. 153134–153157, 2020.

[11] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed Deep Neural Networks Over the Cloud, the Edge, and End Devices," *Proc. IEEE ICDCS*, pp. 328–339, 2017.

[12] R. Roman, J. Lopez, and M. Mambo, "Mobile Edge Computing, Fog et al.: A Survey and Analysis of Security Threats and Challenges," *Future Generation Computer Systems*, vol. 78, pp. 680–698, 2018.

[13] L. Deng and D. Yu, "Deep Learning: Methods and Applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.