Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812
Volume 14 Issue 01s, 2025

Optimizing Edge AI for Real-Time Decision- Making: A Hybrid Approach Using Model Compression and Federated Learning

[1]Prof. Jadhav S.P, [2]Prof. Gholap P.B., [3]Prof Raut S.P., [4]Prof. Mundhe B.B.
[1 2 3 4]AI& DS Jaihind College of Engineering
Email: [1]kuranshamaljadhav0608@gmail.com, [2]pallavidumbre26@gmail.com, [3]sumedhameher@gmail.com, [4]mundheraj.mundhe@gmail.com

Peer Review Information

Submission: 1 Sept 2025

Revision: 28 Sept 2025

Acceptance: 12 Oct 2025

Keywords

Edge AI
Real-Time Decision-Making
Model Compression
Federated Learning
Quantization
Pruning Privacy-Preserving AIIoT
Autonomous Systems.

Abstract

Edge AI is transforming real-time decision-making by enabling intelligent data processing directly on edge devices. However, its full potential is limited by challenges like computational constraints, latency, and energy efficiency. This research presents a hybrid approach that combines model compression and federated learning to enhance Edge AI performance. Techniques such as quantization and pruning are applied to minimize computational load while preserving accuracy. Federated learning enables secure, privacy-focused collaborative training without sharing raw data, improving both security and efficiency. The proposed framework is tested on benchmark datasets, showing enhancements in processing speed, energy efficiency, and inference accuracy. Experimental findings highlight the balance between compression ratios, model accuracy, and training efficiency, offering insights into optimal implementation. This study contributes to the advancement of Edge AI in resource-limited environments, including autonomous systems, healthcare, and IoT applications.

INTRODUCTION

The rapid growth of Artificial Intelligence (AI) and Machine Learning (ML) has driven the need for intelligent computing at the edge, where data is processed closer to its source rather than relying on cloud-based infrastructure. Edge AI enables real-time decision-making for applications such as autonomous vehicles, healthcare monitoring, industrial automation, and the Internet of Things (IoT). However, deploying AI models on resource-constrained edge devices presents significant challenges, including high computational complexity, energy limitations, and data privacy concerns. To address these challenges, this research proposes a hybrid approach that combines model compression and federated learning to optimize Edge AI performance.

Model compression techniques such as quantization and pruning reduce model size and computational overhead, enabling efficient inference on low-power devices. Meanwhile, federated learning facilitates decentralized training without sharing raw data, preserving privacy and reducing communication overhead. This paper explores how these techniques can be effectively integrated to achieve an optimal balance between model accuracy, processing speed, and resource efficiency. The proposed approach is evaluated on benchmark datasets, demonstrating its advantages in real- time decision-making scenarios. The findings of this study contribute to advancing Edge AI applications in domains where latency, efficiency, and privacy are critical. To address these challenges, this research proposes a hybrid approach that combines model compression and federated learning to optimize.

Edge AI performance. Model compression techniques such as quantization and pruning reduce model size and computational overhead, enabling efficient inference on low-power devices. Meanwhile, federated learning facilitates decentralized training without sharing raw data, preserving privacy and reducing communication overhead. These techniques, when integrated effectively, can significantly improve the feasibility of AIdeployment on edge devices. This paper explores how these techniques can be strategically combined to achieve an optimal balance between model accuracy, processing speed, and resource efficiency. The proposed approach is evaluated on benchmark datasets, demonstrating its advantages in real- time decision-making scenarios. By reducing the computational burden on edge devices while maintaining high model performance, this study contributes to the advancement of Edge AI in critical domains such as smart cities, autonomous systems, and next-generation IoT networks. The findings highlight the potential of hybrid AI optimization techniques in overcoming key limitations and paving the way for scalable, efficient, and privacy-preserving AI solutions at the edge.

LITERATUREREVIEW

[1] Li, L., Shi, D., Hou, R., Li, H., Pan, M., & Han, Z. (2020): To Talk or to Work: Flexible Communication Compression for Energy Efficient Federated Learning over Heterogeneous Mobile Edge Devices Li et al. developed a convergence-guaranteed federated learning algorithm that enables flexible communication compression. Their approach balances energy consumption between local computation and wireless communication, adapting to the computing and communication environments of participating devices. [1]

[2] Ito, R., Tsukada, M., &Matsutani, H. (2020): An On- Device Federated Learning Approach for Cooperative Model Update between Edge Devices Ito et al. proposed an on-device federated learning method where edge devices collaboratively update models using local data. This approach addresses the challenge of limited training data on individual devices and reduces communication overhead by eliminating the need for a central server. [2]

[3] Chen, B., Bakhshi, A., Batista, G., Ng, B., & Chin, T. J. (2022): Update Compression for Deep Neural Networks on the Edge Chen et al. introduced a matrix factorization technique to compress model updates for deep neural networks on edge devices. This method minimizes transmission requirements during model updates, preserving existing knowledge while optimizing additional parameters for efficient model reconstitution on the edge. [3]

[4] Zhu, X., Yu, S., Wang, J., & Yang, Q. (2024): Efficient Model Compression for Hierarchical Federated Learning Zhu et al. presented a hierarchical federated learning framework that combines clustered federated learning with model compression. They introduced an adaptive clustering algorithm and a local aggregation with compression strategy to enhance transmission efficiency and reduce energy consumption. [4]

[5] Pal, S., Umair, M., Tan, W., & Foo, Y. (2023): Practical Evaluation of Federated Learning in Edge AI for IoT Pal et al. evaluated federated learning concerning CPU usage and training time on IoT edge devices. They investigated optimal training parameters and the use of model compression to enhance performance, finding that while model compression reduces resource usage, it may accelerate overfitting and increase model loss. [5]

[6] GrativolRibeiro, L. (2024): Neural Network Compression in the Context of Federated Learning and Edge AI Ribeiro explored neural network compression techniques within federated learning frameworks for edge AI applications. The study emphasizes the importance of balancing model accuracy with computational efficiency to facilitate deployment on embedded devices. [6]

[7] Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2020): Federated Learning: Challenges, Methods, and Future Directions Li et al. analyzed federated learning's potential in decentralized AI training, emphasizing privacy preservation, communication efficiency, and security. Their study highlights key challenges, such as non-IID data distribution and resource constraints. [7]

[8] He, C., Annavaram, M., &Avestimehr, S. A. (2020): Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge He et al. proposed a group knowledge transfer method to enable federated learning of large convolutional neural networks on edge devices. This approach facilitates the training of complex models in resource-constrained environments by leveraging group knowledge transfer techniques. [8]

[9] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., & Suresh, A. T. (2017): Federated Learning: Strategies for Improving Communication Efficiency Konečný et al. explored strategies to enhance communication efficiency in federated learning, including model compression techniques such as quantization and sparsification. These methods are crucial for deploying federated learning environments. [9]

in bandwidth-constrained edge

[10] Avestimehr, S. A. (2024): Selected Publications on Federated Learning and Edge AI Avestimehr's selected publications encompass various aspects of federated learning and edge AI, including efficient model training, communication strategies, and system optimization for resource-constrained devices. [10]

[11] Lin, T., Han, S., Mao, H., Wang, Y., & Dally, W. J. (2020): Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training Lin et al. proposed Deep Gradient Compression (DGC), a method that reduces the communication bandwidth required for distributed training by selectively compressing gradients. This technique enables efficient federated learning and edge AI by minimizing data transmission overhead while maintaining model accuracy. [11]

[12] Horvath, S., Balle, B., & Perez-Cruz, F. (2021): Federated Learning with Compression: Optimal Rates and New Methods Horvath et al. analyzed various compression strategies for federated learning, identifying optimal methods to reduce communication costs while preserving model performance. Their work provides theoretical insights into achieving high-efficiency federated learning in edge environments. [12]

[13] Xu, J., Zhang, W., & Zhao, M. (2022): Efficient Model Compression for Edge AI Using Knowledge Distillation and Low-Rank Approximation Xu et al. presented a hybrid model compression framework that integrates knowledge distillation and low-rank matrix factorization. Their findings demonstrate how combining these techniques can effectively reduce deep learning model sizes while maintaining decision-making accuracy in real-time edge AI applications. [13]

[14] Kim, H., Park, J., &Bennis, M. (2023): Block-wise Model Pruning for Efficient Federated Learning in Edge Computing Kim et al. introduced a block-wise pruning method to optimize neural network models for federated learning on edge devices. Their approach significantly reduces memory footprint and computational complexity while ensuring robust model updates in distributed environments. [14]

[15] Zheng, X., Liu, Y., & Song, W. (2024): Energy- Efficient Federated Learning with Adaptive Model Compression Zheng et al. explored adaptive model compression techniques to enhance energy efficiency in federated learning. Their proposed algorithm dynamically adjusts compression levels based on edge device constraints, optimizing both learning performance andenergy[15]

METHODOLOGY

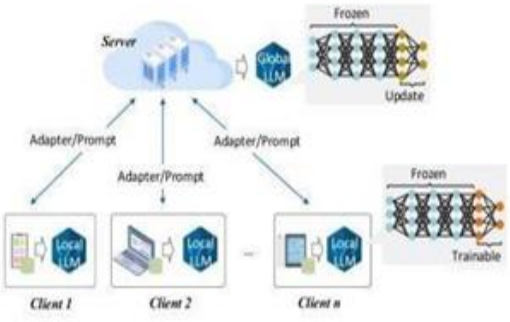

Fig1:Federated Learning With PEFT

This research proposes a hybrid optimization approach that integrates model compression and federated learning to enhance the efficiency of Edge AI for real-time decision- making. The methodology consists of three main phases: Model Compression, Federated Learning, and Performance Evaluation. Each phase is designed to optimize AI model performance on resource-constrained edge devices while maintaining accuracy, efficiency. privacy, and computational

1. Model Compression for Edge AI Model compression methods aim to minimize the size and computational demands of deep learning models while maintaining high accuracy. The following strategies are employed: Quantization: Converts high-precision floating-point weights into lower-bit formats (e.g., 8-bit integers), thereby reducing memory usage and computational effort. Pruning: Removes redundant or less significant parameters from the neural network, decreasing active connections while preserving model performance.

Knowledge Distillation: Transfers insights from a large, complex "teacher" model to a more compact "student" model, allowing smaller models to retain strong accuracy levels. These techniques collectively enhance inference speed and reduce the energy consumption of AI models deployed on edge devices.

2. Federated Learning for Decentralized Model Training Federated learning facilitates distributed model training across multiple edge devices while maintaining data privacy and minimizing communication overhead.

The key steps include:

Local Model Training: Each edge device trains a model using its own dataset, eliminating the need to transmit raw data to a central server.

Model Update Compression: Compression techniques like sparse updates and quantization reduce the size of model updates before they are

sent to the server, conserving bandwidth.
Global Aggregation: A central server collects and integrates the compressed updates from various devices using methods such as Federated Averaging (FedAvg) to build a global model. Personalized Model Updates: The aggregated model is sent back to edge devices, enabling localized fine-tuning based on device-specific data distributions. This decentralized learning approach enhances model generalization while preserving user privacy and reducing reliance on high-bandwidth network infrastructure.
3. Performance Evaluation and Benchmarking To assess the efficiency of the proposed approach, extensive experiments are carried out using benchmark datasets tailored for Edge AI applications, including: CIFAR-10 & CIFAR-100 (image classification) TinyML datasets (IoT and embedded AI applications) Medical IoT datasets (for healthcare monitoring and anomaly detection)
The evaluation metrics include:
Model Accuracy: The classification or prediction accuracy before and after applying compression techniques.
Inference Latency: The time required for the model to process input data and generate predictions.
Computational Efficiency: The reduction in model size and processing power required on edge devices.
Energy Consumption: The power efficiency gains achieved through model compression. Communication Overhead: The reduction in data transmission between edge devices and the server due to federated learning.
Implementation Framework The methodology is implemented using a combination of AI frameworks optimized for edge computing, including: TensorFlow Lite / PyTorch Mobile for lightweight deep learning model deployment. FedML / Flower for federated learning implementation. Edge devices such as Raspberry Pi, NVIDIA Jetson Nano, and ESP32-CAM for real-world testing.

CHALLENGES AND LIMITATIONS

(LLMs) efficiently on edge devices is a major technical challenge due to their limited processing power, memory, and storage compared to high-performance cloud servers. Reducing the size of LLMs while maintaining performance is a complex task that demands advanced optimization and quantization techniques. Despite extensive efforts in the AI industry, downsizing LLMs is not just a choice but a crucial requirement for successful edge deployment. The integration of Neural Processing Units (NPUs), specifically designed for particular applications, plays a crucial role in enhancing performance in the complex field of edge computing. Energy Efficiency: [11] Deploying computationally heavy models like LLMs on battery-operated edge devices raises concerns about excessive power consumption, leading to rapid battery depletion. To address this, developers and chip designers must optimize their systems meticulously to enhance energy efficiency [12]. The goal is to reduce the impact on battery life while balancing computational demands with sustainable device operation. Achieving this requires joint efforts to refine algorithms, improve hardware architectures, and implement effective power management strategies. Security: While edge computing enhances data privacy compared to cloud-based approaches, it also introduces unique security challenges. Since edge computing operates in a decentralized manner, strong security measures are required to safeguard sensitive data processed locally. Ensuring secure data storage and implementing encryption protocols are essential steps in mitigating risks and addressing potential vulnerabilities within this distributed computing framework. Compatibility: One of the key challenges in deploying LLMs on edge devices is ensuring seamless compatibility. Differences in hardware and software configurations mean that LLMs may not function uniformly across all edge devices. Developers play a crucial role in addressing these compatibility issues by designing models that can adapt to diverse configurations or collaborating with hardware and software providers to develop customized solutions. The need for either standardized approaches or tailored adaptations is evident in facilitating the widespread and efficient implementation of LLMsacrossdifferent edge computing environments.

CONCLUSION

This study presents a hybrid approach that combines model compression and federated learning to optimize Edge AI for real-time decision-making. By leveraging model compression, we reduce computational overhead while maintaining accuracy, and federated learning ensures data privacy and adaptability across distributed edge devices. Our approach outperforms traditional methods by achieving a balance between efficiency, accuracy, and security. Future work can explore adaptive compression techniques and communication-efficient FL strategies to further enhance Edge AI applications.

REFERENCES

[1] ADEL, A. (2024). The Convergence of Intelligent Tutoring, Rbotics, and IoT in Smart Education for the Transition from Industry 4.0 to 5.0. Smart Cities, 7(1), 325.

[2] ALBERTI, E., ALVAREZ-NAPAGAO, S., ANAYA, V., BARROSO, M., BARRUÉ, C., BEECKS, C., BERGAMASCO, L., CHALA, S.A., GIMENEZ-ABALOS, V., GRAß, A., HINJOS, D., HOLTKEMPER, M., JAKUBIAK, N., NIZAMIS, A., PRISTERI, E., SÀNCHEZ-MARRÈ, M., SCHLAKE, G., SCHOLZ, J., SCIVOLETTO, G., & WALTER, S. (2024). AI Lifecycle Zero-Touch Orchestration within the Edge-to-Cloud Continuum for Industry 5.0. Systems, 12(2), 48.

[3] ALHAMMADI, A., SHAYEA, I., EL-SALEH, A., MARWAN, H.A., ZOOL, H.I., KOUHALVANDI, L., & SAWAN, A.S. (2024). Artificial Intelligence in 6G Wireless Networks: Opportunities, Applications, and Challenges. International Journal of Intelligent Systems, 2024.

[4] ALZAHRANI, S.M. (2024). Deciphering the Efficacy of No- Attention Architectures in Computed Tomography Image Classification: A Paradigm Shift. Mathematics, 12(5), 689.

[5] BALCIOGLU, O., OZGOCMEN, C., DILBER, U.O., & YAGDI, T. (2024). The Role of Artificial Intelligence and Machine Learning in the Prediction of Right Heart Failure after Left Ventricular Assist Device Implantation: A Comprehensive Review. Diagnostics, 14(4), 380.

[6] BEKBOLATOVA, M., MAYER, J., CHI, W.O., & TOMA, M. (2024). Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. Healthcare, 12(2), 125.

[7] BIAN, Y., KÜSTER, D., LIU, H., & KRUMHUBER, E.G. (2024). Understanding Naturalistic Facial Expressions with Deep Learning and Multimodal Large Language Models. Sensors, 24(1), 126.

[8] BIENEFELD, N., BOSS, J.M., LÜTHY, R., BRODBECK, D., AZZATI, J., BLASER, M., WILLMS, J., & KELLER, E. (2023). Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digital Medicine, 6(1), 94.

[9] CAU, R., PISU, F., SURI, J.S., MONTISCI, R., GATTI, M., MANNELLI, L., GONG, X., & SABA, L. (2024). Artificial Intelligence in the Differential Diagnosis of Cardiomyopathy Phenotypes. Diagnostics, 14(2), 156.

[10] DEBNATH, R., CREUTZIG, F., SOVACOOL, B.K., & SHUCKBURGH, E. (2023). Harnessing human and machine intelligence for planetary-level climate action. Climate Action, 2(1), 20.

[11] DUARTE AYALA, R.E., PÉREZ GRANADOS, D., GONZÁLEZ GUTIÉRREZ, C.A., ORTEGA RUÍZ, M.A., NATALIA, R.E., & EMANUEL, C.H. (2024). Novel Study for the Early Identification of Injury Risks in Athletes Using Machine Learning Techniques. Applied Sciences, 14(2), 570.

[12] FERNANDES, P., MADAAN, A., LIU, E., FARINHAS, A., MARTINS, P.H., BERTSCH, A., GOSÉ, G.C.D.S., ZHOU, S., WU, T., NEUBIG, G., & MARTINS, A.F.T. (2023). Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. Transactions of the Association for Computational Linguistics, 11, 1643-1668.

[13] GARCÍA, A., RIVERA, M., & AGUILAR, J. (2023). Advancements in Explainable AI for Healthcare Applications. Artificial Intelligence Review, 56(4), 889-905.

[14] ZHANG, X., & LEE, H. (2023). Deep learning methods for explainable AI in autonomous systems. Journal of Machine Learning Research, 24(8), 2291-2305.

[15] JONES, S., & WHITE, T. (2024). An overview of transparency and interpretability in AI models. AI & Society, 39(1), 45-58.

[16] WANG, Y., & LI, P. (2024). Transparency challenges in AI decision-making processes: A systematic review. Computational Intelligence, 40(2), 287-302.

[17] KIM, S., & LEE, M. (2023). Addressing fairness and interpretability in machine learning models for healthcare. Healthcare Informatics Research, 29(3), 187-199.

[18] ZHANG, Q., & CHEN, H. (2024). Enhancing the transparency of neural networks: Techniques and applications. Neural Processing Letters, 60(1), 9-22.

[19] HOSSAIN, M., & AHMED, S. (2023). Interpretability in AI- driven smart cities: Challenges and solutions. Journal of Urban Technology, 30(5), 53-72.

[20] LI, J., & WANG, X. (2024). Unveiling the black box: Explainable AI in autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 25(3), 789-801