ISSN : 2455-9679
SJIF Impact Factor : 6.190

IJRTSM
INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"DEVELOPMENT OF ENERGY EFFICIENT FEDERATED LEARNING MODELS FOR REDUCING THE COMPUTATIONAL BURDEN ON EDGE DEVICES: A LITERATURE REVIEW"

Nidhi Paliwal [1], ***Amlesh Singh*** [2]

[1] *Research Scholar, Department of Computer Science and Application, Rabindranath Tagore University, Raisen, Madhya Pradesh, India*

[2] *Assistant Professor, Department of Computer Science and Application, Rabindranath Tagore University, Raisen, Madhya Pradesh, India*

nidhipaliwal237@gmail.com
amlesh.singh@rntu.ac.in

Corresponding Author: amlesh.singh@rntu.ac.in

ABSTRACT

The excess of Internet of Things (IoT) tools and edge intelligence has accelerated the adoption of Federated Learning (FL) as a secrecy sustaining paradigm for distributed model training. While FL eliminates the requirement to integrate raw data, its practical deployment on resource-constrained edge devices is obstructed by significant computational, energy, and communication overheads. This literature review investigates recent advances in the development of energy-efficient FL models, with an emphasis on techniques that minimize device-level burdens without compromising learning accuracy. Key strategies identified in the literature include gradient compression (quantization, pruning, and sparsification), hierarchical and clustered aggregation, adaptive client participation, and reinforcement learning-based resource allocation. Comparative studies demonstrate that these methods can substantially reduce training time, communication costs, and device-level energy consumption, while maintaining competitive model performance. However, trade-offs between accuracy and efficiency, system heterogeneity, and dynamic network conditions remain open issues. This review not only synthesizes the state of the art but also outlines future research opportunities to design scalable, sustainable, and practical FL frameworks for edge computing environments.

Key Words: *Quantization, Pruning, Sparsification, Federated Learning, Internet of Things.*

I. INTRODUCTION

The quick growth of the IoT, mobile computing, and edge intelligence has directed to an exponential enhance in the generation of distributed data across heterogeneous devices. Traditional centralized machine learning concepts, which depend on gathering raw data in a centralized server for training, are increasingly becoming impracticable because of secrecy concerns, high communication overhead, and significant energy consumption [10]. The FL has appeared as a auspicious concept to address these issues by allowing collaborative model training throughout distributed devices deprived of distributing raw data. Although its benefits in sustaining secrecy and decreasing data transfer costs, FL introduces new challenges, particularly in terms of computational and energy efficiency when deployed on resource-constrained edge devices [1, 7].

Edge tools like smartphones, IoT sensors, and embedded systems are typically characterized by limited processing

ISSN : 2455-9679
SJIF Impact Factor : 6.190

power, restricted memory, battery constraints, and varying network conditions. Executing multiple local training iterations and transmitting large model updates to the server can impose a substantial computational burden on these devices, leading to reduced participation, degraded performance, and premature device failures in real-world FL systems. Consequently, the development of energy-efficient FL models has become a critical research focus to ensure scalability, sustainability, and practical deployment of FL at the network edge [4, 16].

Several strategies have been implemented in the literature to reduce the computational and communication burden in FL while maintaining or even improving model accuracy. These include gradient compression techniques such as quantization, pruning, and sparsification; hierarchical and clustered aggregation methods that reduce redundant communication; adaptive device selection mechanisms to involve only the most suitable participants; and reinforcement learning-based resource allocation approaches that optimize training under dynamic conditions [8]. Moreover, hybrid approaches integrating edge-cloud collaboration and model optimization have demonstrated significant promise in balancing performance with energy efficiency [3, 7].

This literature review aims to offer a comparative review of latest advancements in energy effective FL, aiming on techniques that reduce computational complexity and communication costs on edge devices. The review highlights the trade-offs among model precision, energy expenses, communication efficiency, and system scalability. By critically analysing state-of-the-art models like hierarchical FL, quantization-based approaches, device selection frameworks, and reinforcement learning-driven optimizations, this study identifies key research gaps and opportunities for designing sustainable and practical FL systems [12, 13].

Ultimately, the development of energy-efficient FL models is crucial for enabling large-scale adoption of intelligent services in domains like smart healthcare, autonomous devices, industrial IoT, and next-generation wireless networks. This review sets the stage for future research directions in building optimized FL frameworks that ensure high accuracy with minimal energy and computational burden on edge devices, thus aligning with the broader vision of green AI and sustainable edge intelligence [5, 6].

II. Literature Review

The EaMC-FL model [1] introduces an energy-aware federated learning strategy that reduces transmission costs in wireless networks by selecting representative edge devices through a multi-criteria evaluation process. Edge nodes are grouped according to parameter matching, and a subset of them is chosen for training by balancing performance, energy usage, and battery lifetime. Experimental outcomes confirm that this strategy significantly lowers energy consumption while keeping model quality, preparing it fit for energy obliged IoT ecosystems.

The EAFL+ framework [2] tackles the high computational and energy costs in federated learning by leveraging a cloud–edge–terminal collaborative architecture. It dynamically offloads computation to the most optimal resource tier, balancing power consumption and training efficiency. By managing resource diversity and selecting appropriate offloading targets, EAFL+ achieves higher accuracy, faster convergence, and zero client dropouts, outperforming earlier approaches such as EAFL and Oort by up to 24% in accuracy gains and 9% in convergence improvements.

The HED-FL framework [3] integrates hierarchical aggregation and adaptive learning in federated learning to enable energy-efficient edge intelligence. By structuring the FL process across multiple edge layers, it reduces redundant transmissions and computational overhead while maintaining strong model accuracy. Evaluations demonstrate that HED-FL effectively balances accuracy, loss minimization, and energy savings, thereby addressing scalability and sustainability issues in IoT-driven edge learning.

The FedGreen framework [4] enhances FL in mobile edge computing using fine-grained gradient compression, minimizing energy consumption without compromising accuracy. It applies gradient reduction at the device level and element-wise gathering at the server, optimizing the trade-off among precision and energy effectiveness. Experiments confirm that FedGreen reduces device energy consumption by at least 32% while meeting accuracy requirements, outperforming baseline schemes significantly.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

The AutoFL approach [5] applies reinforcement learning to address system heterogeneity and non-IID data challenges in edge FL. By dynamically selecting participating devices and optimizing execution targets, AutoFL accelerates convergence and enhances energy efficiency under stochastic conditions. Results show AutoFL offers 3.6× faster convergence and 4.7–5.2× higher energy efficiency compared to baseline systems, making it highly effective for real-world deployments.

A delay-efficient FL framework [6] is proposed to mitigate communication and computation delays in FL over mobile edge devices. By formulating the compromise among wireless communication ("to talk") and local computation ("to work") as an optimization problem, the method minimizes total training time. Simulation results show the approach reduces communication rounds and achieves faster convergence despite network uncertainties, enabling more efficient deployment of FL in mobile edge environments.

This work [7] introduces a joint design of weight quantization and wireless transmission optimization for energy-efficient federated learning on heterogeneous mobile devices. By modeling the problem as a mixed-integer programming optimization, it minimizes total energy consumption while guaranteeing accuracy and latency requirements. An iterative algorithm determines optimal quantization levels and bandwidth allocation, achieving substantial energy savings and training efficiency compared to conventional FL designs.

A sparsification and optimization framework [8] for federated learning is proposed to decrease the computational and communication costs on resource obliged devices. By introducing sparse neural networks and enhancing gradient descent algorithms, the system lowers both computation and transmission energy. Combined with optimization of bandwidth, power, and learning parameters, the method achieves significant reductions in energy consumption while maintaining model accuracy, outperforming traditional fully-connected FL systems.

This study [9] proposes a deep reinforcement learning (DRL)-based offloading and resource allocation approach for Industrial Internet mobile edge computing. Terminal devices can devolve tasks either fully or partially to edge servers, reducing local energy consumption. By optimizing offload ratios, transfer strength, and computing frequencies, the DRL-enhanced FL system achieves rapid convergence and significantly reduces total energy consumption compared to baseline methods, ensuring efficient and sustainable industrial edge intelligence.

The Cognitive Energy Management Scheme (CEMS) [10] addresses edge device exhaustion and inefficient offloading in federated learning systems. Using state learning, CEMS dynamically switches between computation and offloading states to balance energy usage. It improves computing rate and energy efficiency by over 7%**,** while reducing devolve ratio, scheduling collapses, and evaluation time by up to 15%**.** This scheme outperforms conventional wireless edge strategies by preventing premature device exhaustion and optimizing energy usage.

An Alternative Direction Algorithm (ADA) [11] is implemented for improving bandwidth distribution, CPU frequency, and transfer strength in IoT edge intelligence-based federated learning. By addressing bandwidth allocation in closed form and iteratively refining computational resources, the method reduces device energy consumption without significantly extending FL training time. Simulation results confirm that ADA adapts well to varying system conditions, balancing energy efficiency and accuracy effectively for large-scale IoT deployments.

The FL-TD3 framework [12] applies twin-hindered deep deterministic strategy incline reinforcement learning to improve energy efficiency and accuracy trade-offs in mobile edge federated learning. It formulates energy-efficient scheduling as a continuous optimization problem, enabling devices to maximize learning accuracy relative to energy usage. Results demonstrate that FL-TD3 significantly improves performance associated to state-of-the-art techniques, offering robust energy savings while maintaining high accuracy.

The E2DS device selection framework [13] addresses the limitations of existing FEL systems by jointly optimizing energy consumption and data diversity in device participation. Unlike prior works focusing mainly on time and data amount, E2DS ensures energy-efficient operation while minimizing client dropouts. With a novel algorithm reducing time complexity, experiments demonstrate superior performance in training time and model robustness compared to classical federated learning methods.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

The SpFedRec framework [14] introduces a split learning and cloud-edge collaboration approach for large-scale federated recommendation systems. By migrating section of the model to the cloud and compressing item data, it reduces device-side computation and communication overhead. Additionally, privacy is enhanced via a multi-party orbitual secret-sharing scheme. Experiments show SpFedRec decreases computation time by 23% and communication cost by 49%**,** while maintaining competitive accuracy with state-of-the-art FL recommender systems.

A unified anomaly detection framework [15] for wireless sensor networks combines Ensemble FL (EFL) with energy-efficient online learning (OAD-EE) and cloud integration. EFL enhances revealing precision while OAD-EE conserves energy and improves revealing speed. Together, they achieve both high accuracy and low energy consumption, enabling scalable and real-time anomaly exposure in energy-constrained WSNs, making the approach ideal for industrial applications requiring reliability and sustainability.

A model pruning-based FL framework [16] is proposed to decrease computation and communication costs on resource obliged edge devices. By dynamically pruning neurons or kernels and reusing parameters, it maintains accuracy while reducing device workload. Extensive experiments on MNIST and CIFAR-10 show computation reductions of 63–72% and memory savings of 59–72%**,** while achieving accuracy surpassing baseline methods. This makes the approach well-suited for bandwidth-limited and low-power edge environments.

Table 1 describes the comprehensive revies of prior works on energy efficient federated learning approaches.

Table 1: Comprehensive Review

Paper	Proposed Methodology	Performance Factors	Advantages	Limitations
A. A. Al-Saedi et al., 2021 [1]	Energy effective multi-criteria FL model for edge computing	Energy consumption, model accuracy, computation time	Energy efficiency, reduced latency	Complex implementation, potential model degradation
A. Arouj and A. M. Abdelmoniem, 2024 [2]	Collaborative computing approach for energy-aware federated learning	Energy consumption, data transfer efficiency, computation load	Enhanced energy efficiency, improved collaboration	May require high initial resource allocation, possible scalability issues
F. D. Rango et al., 2023 [3]	HED-FL: Hierarchical, energy-efficient, and dynamic approach	Hierarchical model efficiency, energy consumption, communication overhead	Adaptive energy management, scalable	Increased complexity, communication delays in hierarchy
P. Li et al., 2021 [4]	FedGreen: Fine-grained gradient compression in federated learning	Communication efficiency, energy consumption, model convergence	Reduced communication cost, energy-efficient	Potential loss of accuracy due to compression, complex gradient management
Y. G. Kim and C. J. Wu, 2021 [5]	AutoFL: Heterogeneity-aware energy-efficient federated learning	Device heterogeneity, energy consumption, training time	Adaptability to device capabilities, energy-efficient	Requires extensive device profiling, potential compatibility issues

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Paper	Proposed Methodology	Performance Factors	Advantages	Limitations
P. Prakash et al., 2021 [6]	Delay aware FL over mobile edge devices	Training delay, energy consumption, model accuracy	Reduced delay, energy-efficient	Limited by network conditions, potential model accuracy trade-offs
R. Chen et al., 2023 [7]	Combine Model of weight quantization and wireless transfer for energy-efficient FL	Weight quantization impact, energy efficiency, communication delay	Reduced energy consumption, optimized communication	Possible loss in model accuracy, complexity in joint optimization
L. Lei et al., 2022 [8]	Sparsification and optimization for energy proficient FL	Sparsification efficiency, energy consumption, model performance	Reduced communication overhead, energy-efficient	Potential sparsity-induced accuracy degradation, optimization complexity
X. Li et al., 2023 [9]	Federated deep reinforcement learning	Resource allocation efficiency, energy consumption, model convergence	Optimal resource utilization, energy-efficient	High computational complexity, slow convergence
V. K. Kaliappan et al., 2022 [10]	Cognitive energy management for energy-efficient offloading in edge computing	Offloading efficiency, energy consumption, task completion time	Enhanced energy management, efficient task offloading	Complex cognitive management, dependency on accurate predictions
A. Salh et al., 2023 [11]	Resource allocation for Green IoT edge devices	Resource utilization, energy efficiency, learning speed	Green IoT support, energy-efficient	Complex resource allocation, potential bottlenecks in scalability
J. Zheng et al., 2023 [12]	FL with deep reinforcement learning for resource allocation	Resource allocation efficiency, energy consumption, model adaptability	Dynamic resource allocation, energy-efficient	High computational demand, potential delay in learning convergence
C. Peng et al., 2021 [13]	Energy aware device choice in FL	Device selection efficiency, energy consumption, learning accuracy	Improved energy efficiency, optimal device utilization	Possible suboptimal learning accuracy, selection overhead
J. Qin et al., 2023 [14]	Split FL with edge-cloud integration for privacy-preserving recommendations	Privacy preservation, energy efficiency, model scalability	Enhanced privacy, energy-efficient, scalable	Split learning complexity, potential latency issues
S. Gayathri and D. Surendran,	Unified collaborative FL	Anomaly detection accuracy, energy	Improved anomaly detection, energy-	Computational complexity, potential

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Paper	Proposed Methodology	Performance Factors	Advantages	Limitations
2024 [15]	with cloud computing	efficiency, computation load	efficient	cloud dependency
T. Wu et al., 2023 [16]	Model pruning for effective FL on resource obliged devices	Pruning effectiveness, energy consumption, model accuracy	Reduced model size, energy-efficient, suitable for resource-constrained devices	Possible accuracy loss due to pruning, requires careful tuning

III. Research Gap

Here are the overall research gaps identified across the provided papers:

1. The prior studies focus on energy-aware federated learning models but lacks extensive experimentation on diverse real-world datasets and edge devices. There's a need for more comprehensive evaluations in dynamic and heterogeneous environments.

2. The hierarchical approach is promising, but the prior works don't address the scalability challenges in large-scale networks. The impact of latency and communication overheads in different network topologies is also underexplored.

3. The method's effectiveness across various learning tasks and datasets remains unclear. Although the method adapts to device heterogeneity, it requires extensive profiling, which might not be practical in all settings.

4. The deep reinforcement learning strategy for resource allocation is promising, but the model's high computational complexity and slow convergence are potential barriers for real-time applications. There's also a need for more robust testing in industrial IoT scenarios.

5. The cognitive energy management scheme is complex and may require high computational resources, which can be impractical in real-world edge environments. Further work could focus on simplifying the cognitive processes or developing more efficient algorithms.

6. The prior works focus on resource allocation for green IoT but do not fully consider the potential security vulnerabilities in federated learning. Future research could explore secure resource allocation methods that are also energy-efficient.

7. While the deep reinforcement learning approach is dynamic, its maximum computational need might limit its ability in resource obliged environments. More efficient algorithms with lower computational requirements could be explored.

8. The split federated learning model is privacy-preserving but may suffer from latency issues due to edge-cloud integration. Future work could focus on optimizing the communication protocols to reduce latency without compromising privacy.

IV. Conclusion

Federated Learning represents a transformative step toward distributed and privacy-preserving AI, yet its effectiveness is constrained by the minimum computational and energy resources of edge devices. This review highlights the necessity of energy-efficient FL model that provide a balance between accuracy, communication efficiency, and system scalability. Existing literature demonstrates that approaches such as quantization, pruning, sparsification, and hierarchical aggregation can significantly reduce computational and communication overhead, while adaptive client

participation and intelligent resource management further enhance efficiency. Nonetheless, challenges persist, including handling heterogeneous device capabilities, ensuring fairness in client participation, mitigating accuracy loss from aggressive compression, and adapting to dynamic edge environments. Addressing these gaps requires the integration of hybrid edge-cloud solutions, adaptive optimization strategies, and cross-layer design approaches aligned with green AI principles. Ultimately, the development of energy-efficient FL models is essential for enabling widespread deployment of intelligent services across domains like healthcare, smart cities, and autonomous systems, ensuring sustainability and scalability in next-generation edge intelligence.

REFERENCES

[1] A. A. Al-Saedi, E. Casalicchio and V. Boeva, "An Energy-aware Multi-Criteria Federated Learning Model for Edge Computing", Proceedings -2021 International Conference on Future Internet of Things and Cloud, FiCloud, IEEE, pp. 134-143, 2021.

[2] A. Arouj and A. M. Abdelmoniem, "Towards Energy Aware Federated Learning via Collaborative Computing Approach", Computer Communications, 221, pp. 131-141, 2024.

[3] F. D. Rango, A. Guerrieri, P. Raimondo and G. Spezzano, "HED-FL: A Hierarchical, Energy efficient and Dynamic approach for Edge Federated Learning", Pervasive and Mobile Computing, Elsevier, 92, pp. 1-21, 2023.

[4] P. Li, X. Huang, M. Pan and R. Yu, "FedGreen: Federated Learning with Fine-Grained Gradient compression for Green Mobile Edge Computing", arXiv:2111.06146v1 [cs.LG], pp. 1-6, 2021.

[5] Y. G. Kim and C. J. Wu, "AutoFL: Enabling Heterogeneity-Aware Energy Efficient Federated Learning", Micro`21, ACM, Athens, Greece, pp. 183-198, 2021.

[6] P. Prakash, J. Ding, M. Wu, M. Shu, R. Yu and M. Pan, "To Talk or to Work: Delay Efficient Federated Learning over Mobile Edge Devices", IEEE Xplore, pp. 1-6, 2021.

[7] R. Chen, L. Li, K. Xue, C. Zhang, M. Pan, Y. Fang, "Energy Efficient Federated Learning over Heterogeneous Mobile Devices via Joint Design of Weight Quantization and Wireless Transmission", IEEE Transaction on Mobile Computing, 22 (12), pp. 7451-7465, 2023.

[8] L. Lei, Y. Yuan, Y. Yang, Y. Luo, L. Pu and S. Chatzinotas, "Sparsification and Optimization for Energy Efficient Federated Learning in Wireless Edge Networks", IEEE, pp. 1-8, 2022.

[9] X. Li, J. Zhang and C. Pan, "Federated Deep Reinforcement Learning for Energy Efficient Edge Computing Offloading and Resource Allocation in Industrial Internet", Applied Sciences, MDPI, 13 (6708), pp. 1-24, 2023.

[10] V. K. Kaliappan, A. B. L. Ranganathan, S. Periasamy, P. Thirumalai, T. A. Nguyen, S. Jeon, D. Min and E. Choi, "Energy Efficient Offloading based on Efficient Cognitive Energy Management Scheme in Edge Computing Device with Energy Optimization", Energies, MDPI, 15 (8273), pp. 1-16, 2022.

[11] A. Salh, R. Ngah, L. Audah, K. S. Kim, Q. Abdullah, Y. M. A. Moliki, K. A. Aljaloud and H. N. Talib, "Energy Efficient Federated Learning with Resource Allocation for Green IoT Edge Intelligence in B5G", IEEE Access, 11, pp. 1-15, 2023.

[12] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar and M. Guizani, "Federated Learning for Online Resource Allocation in Mobile Edge Computing: A Deep Reinforcement Learning Approach", In Proceeding of the 2023 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, pp. 1-6, 2023.

[13] C. Peng, Q. Hu, J. Chen, K. Kang, F. Li and X. Zou, "Energy Efficient Device Selection in Federated Edge Learning", International Conference on Computer Communications and Networks (ICCCN), pp. 1-9, 2021.

[14] J. Qin, X. Zhang, B. Liu and J. Qian, "A Split Federated Learning and Edge-Cloud based Efficient and Privacy Preserving Large Scale Item Recommendation Model", Journal of Cloud Computing: Advances, Systems and Application, 12 (57), pp. 1-17, 2023.

[15] S. Gayathri and D. Surendran, "Unified Ensemble Federated Learning with Cloud Computing for Online Anomaly Detection in Energy Efficient Wireless Sensor Networks", Journal of Cloud Computing: Advances, Systems and Applications, 13 (49), pp. 1-21, 2024.

[16] T. Wu, C. Song and P. Zeng, "Efficient Federated Learning on Resource Constrained Edge Devices based on Model Pruning", Complex & Intelligent Systems, 9, 6999-7013, 2023.