Volume 1, Issue 2, October 2025 eISSN: 2734-8119
DOI: https://doi.org/10.5281/ijurd.v1i2.6

International Journal of Unified Research & Development
Journal Homepage: ijurd.com

Federated Learning in Cloud–Edge Computing for Privacy-Preserving and Scalable Resource Management

Shyam Sunder Saini[1]*

[1] *University of Jammu,* Baba Saheb Ambedkar Road, Tawi, Jammu, Jammu and Kashmir 180006, India
Email: sundershyam51@gmail.com

ABSTRACT

The exponential growth of Internet of Things (IoT) devices, real-time applications, and data-driven services has exposed the limitations of centralized cloud computing, particularly in terms of scalability, latency, energy consumption, and data privacy. Cloud–edge computing has emerged as a paradigm shift by bringing computation closer to data sources; however, traditional machine learning approaches still require centralizing sensitive information, creating risks of privacy violations and inefficient use of distributed resources. To address these challenges, this research proposes a hierarchical federated learning (FL) framework for privacy-preserving and scalable resource management in cloud–edge ecosystems. In the proposed architecture, client devices train models locally, edge servers perform intermediate aggregation, and the cloud coordinates global aggregation and orchestration, while communication-efficient strategies such as quantization and sparsification, robust aggregation methods, and differential privacy mechanisms are integrated to optimize bandwidth, energy, and security trade-offs. The framework is implemented using CloudSim Plus, iFogSim2, and TensorFlow Federated, and validated with heterogeneous workloads, including healthcare datasets (MIMIC-III) and smart city sensor/mobility traces. Results demonstrate that hierarchical FL reduces communication overhead by up to 60%, improves latency by 25–35%, and lowers energy consumption by 20–25%, while maintaining accuracy levels close to centralized machine learning. Privacy-preserving mechanisms ensure regulatory compliance and safeguard sensitive information, while robust aggregation enhances resilience against adversarial attacks. Compared to centralized ML and naïve federated learning, the proposed system achieves the most balanced trade-off between efficiency, scalability, accuracy, and confidentiality. Beyond technical contributions, this work offers direct societal benefits: in healthcare, it enables hospitals to collaboratively train diagnostic models without exposing patient data; in smart cities, it supports traffic forecasting, energy optimization, and public safety with low-latency intelligence; and in finance, it strengthens fraud detection while preserving data sovereignty. By aligning with current demands for trustworthy, sustainable, and inclusive digital infrastructures, this study positions federated learning in cloud–edge systems as a key enabler of next-generation intelligent services that are not only efficient but also socially responsible.

Keywords: Federated Learning (FL), Cloud–Edge Computing, Privacy-Preserving Machine Learning, Resource Management, Scalable Intelligent Systems

1. INTRODUCTION

The rapid proliferation of data-intensive applications such as real-time video analytics, smart healthcare monitoring, autonomous driving, financial fraud detection, and immersive Internet of Things (IoT) services has accelerated the reliance on cloud computing as the backbone of digital infrastructure [1]. Cloud computing offers elastic resources, on-demand scalability, and high-performance services, which collectively drive innovation across domains. However, the exponential growth of connected devices and latency-sensitive applications has also highlighted the inherent limitations of centralized cloud infrastructures [2]. Centralized data processing often suffers from network congestion, high latency, privacy risks, and limited responsiveness, especially in scenarios where billions of IoT devices continuously generate massive volumes of data at the network edge. To address these constraints, cloud–edge computing paradigms have emerged, wherein computing, storage, and networking capabilities are extended closer to end-users by leveraging edge servers, fog nodes, and micro data centers. Cloud–edge synergy enables lower latency, localized data processing, and bandwidth savings, thereby enhancing user Quality of Service (QoS) while retaining the computational power of centralized cloud infrastructures [3], [4].

In parallel, the field of machine learning (ML) has transformed how intelligent services are delivered across distributed environments. ML models underpin predictive analytics, recommendation systems, anomaly detection, and decision-making mechanisms that are essential for optimizing cloud–edge resource management. Yet, the traditional centralized approach to

ML training, where raw data is collected from end devices and uploaded to a central server for model training, is increasingly infeasible [5]. Not only does this approach impose significant communication overheads, but it also raises serious concerns regarding data privacy, security, and regulatory compliance. Sensitive domains such as healthcare, finance, and critical infrastructure cannot freely transmit raw data to cloud servers due to legal frameworks such as GDPR, HIPAA, and CCPA. Consequently, there has been a paradigm shift toward privacy-preserving distributed learning approaches [6].

Federated Learning (FL) has emerged as a powerful framework to bridge this gap. In FL, model training occurs collaboratively across distributed devices or edge nodes without requiring the raw data to leave the local environment. Each participating node trains the model locally using its own data and then transmits only the updated model parameters (e.g., gradients) to a central aggregator, typically hosted in the cloud [7]. The aggregator combines these updates into a global model, which is redistributed for further training in iterative rounds. This architecture enables data sovereignty, enhanced privacy, and efficient utilization of distributed resources. FL, when integrated into cloud–edge computing ecosystems, offers a compelling paradigm for scalable and privacy-preserving resource management. Edge devices contribute computational power and localized learning, while the cloud provides orchestration, global aggregation, and long-term storage [8], [9].

Despite its promise, integrating FL into cloud–edge environments presents a series of technical challenges that require rigorous investigation. One central issue is heterogeneity: edge devices and IoT sensors exhibit diverse computational capabilities, network conditions, and data distributions [10]. This non-independent and identically distributed (non-IID) data problem often reduces the accuracy and convergence rate of federated models. Another challenge is communication efficiency, as transmitting model updates from thousands or millions of devices to cloud aggregators can overwhelm networks, especially under bandwidth-constrained or mobile scenarios. Moreover, security vulnerabilities such as poisoning attacks, model inversion, and Byzantine failures threaten the reliability of federated models. Ensuring fault tolerance and energy efficiency also remains critical, since edge devices often operate under limited power budgets [11]. At the same time, cloud–edge resource management introduces its own complexities. Resource allocation, task scheduling, service placement, and workload balancing must be optimized dynamically to achieve performance goals such as minimizing latency, reducing operational costs, and maximizing energy efficiency. Traditional optimization methods struggle to scale under such high-dimensional, dynamic, and uncertain environments [12]. Here, machine learning-driven resource management has shown promise, but conventional centralized ML approaches fall short due to privacy risks and inefficiencies in data transfer. Therefore, federated learning stands out as an enabler of intelligent, scalable, and privacy-preserving resource management strategies. By allowing edge devices to collaboratively learn workload prediction models, anomaly detection systems, or energy-efficient scheduling policies without exposing raw data, FL provides a foundation for robust and adaptive cloud–edge ecosystems [13]. The significance of this integration extends beyond technical optimization to societal and economic implications. In healthcare, federated cloud–edge systems allow hospitals to train diagnostic models on sensitive patient data without violating privacy regulations. In smart cities, edge sensors can collaboratively learn models for traffic prediction, energy consumption forecasting, and public safety monitoring while minimizing cloud dependency [14],[15]. In finance, distributed fraud detection models trained across banks and ATMs preserve confidentiality while combating evolving cyber threats. Thus, federated learning in cloud–edge environments aligns with global priorities for secure, sustainable, and human-centered digital infrastructure [16], [17], and [18].

Existing literature has made strides in exploring federated learning for edge computing and resource management; however, several research gaps remain unaddressed. First, most current frameworks focus on algorithmic improvements for FL without deeply considering the system-level constraints of cloud–edge orchestration. Second, many works emphasize accuracy and privacy but neglect scalability, energy efficiency, and communication cost, which are critical in real-world deployments [19], [20]. Third, limited studies examine the interplay between federated learning, security mechanisms (such as blockchain), and multi-cloud interoperability. Moreover, benchmarking and standardized evaluation metrics for FL in cloud–edge scenarios remain insufficient, hindering reproducibility and comparative analysis [21], [22] and [23].

This manuscript is motivated by the need to systematically address these gaps. Specifically, it explores how federated learning can be leveraged for privacy-preserving and scalable resource management in cloud–edge ecosystems. The objectives of this study are threefold: (i) to analyze the limitations of traditional cloud and edge resource management strategies in terms of privacy, scalability, and efficiency; (ii) to design and evaluate FL-driven approaches for optimizing workload scheduling, service placement, and fault tolerance under heterogeneous environments; and (iii) to propose a framework that ensures both robust privacy protection and sustainable scalability. By doing so, this research aims to contribute a novel, interdisciplinary perspective that integrates machine learning, distributed systems, and cloud–edge orchestration to advance the state of the art in intelligent resource management.

Ultimately, the integration of federated learning into cloud–edge computing is not simply a technological enhancement but a transformative shift toward decentralized, trustworthy, and adaptive digital infrastructures. As the world progresses toward

ubiquitous computing, autonomous systems, and intelligent services, the ability to manage resources securely and efficiently at scale becomes paramount. This study seeks to demonstrate that federated learning-enabled cloud–edge architectures represent a critical step in achieving this vision, paving the way for future innovations in smart healthcare, connected vehicles, sustainable cities, and beyond.

2. LITERATURE REVIEW

Federated Learning (FL) has rapidly emerged as a transformative paradigm for enabling collaborative intelligence across distributed cloud–edge ecosystems while preserving privacy. Unlike traditional centralized learning, FL eliminates the need to transmit raw data to cloud servers by allowing model training to occur locally at the edge, followed by the aggregation of model parameters into a global model. This design is highly relevant to cloud–edge computing where data is generated in massive volumes by heterogeneous IoT devices, healthcare systems, and smart infrastructures. However, achieving scalable, efficient, and privacy-preserving FL across distributed and resource-constrained environments introduces multiple challenges such as communication bottlenecks, device heterogeneity, energy consumption, non-IID data, and adversarial threats. Over the last five years, researchers have proposed a range of solutions, focusing on communication-efficient protocols, hierarchical architectures, adaptive scheduling, robust aggregation, and privacy-aware mechanisms. The following section critically reviews recent works (2020–2025) that explore the interplay of FL and cloud–edge computing for scalable resource management.

A major research thread has focused on communication efficiency, as communication between clients and the cloud aggregator often dominates the cost in FL. Diao et al. proposed HeteroFL, which tackles system heterogeneity by allowing clients with different computational capabilities to train subnetworks of varying sizes, thereby enabling broader participation without forcing weaker devices to drop out [24]. Although designed for heterogeneity, this approach implicitly reduces communication by transmitting smaller models from low-capacity devices. Reisizadeh et al. advanced communication efficiency with FedPAQ, which integrates periodic averaging (allowing multiple local updates before communication) and quantized gradient exchange to cut down uplink bandwidth [25]. The results showed that FedPAQ significantly reduces communication while retaining accuracy comparable to baseline FL methods. In a similar vein, Li et al. introduced FedSparse, which regularizes gradients to enforce sparsity, allowing only important parameter updates to be transmitted [28]. This approach drastically reduces bandwidth without compromising model quality, making it suitable for large-scale cloud–edge deployments where bandwidth is a scarce resource. Chen et al. contributed a theoretical perspective by analyzing generalized communication-efficient strategies that combine compression and sparsification while proving strong convergence guarantees [27]. Collectively, these studies demonstrate that communication efficiency remains a core challenge, with emerging strategies converging on selective communication, compression, and hierarchical aggregation.

Complementing communication efficiency, hierarchical FL architectures have been proposed to better exploit the natural hierarchy in cloud–edge systems. Wang et al. designed a hierarchical FL model where intermediate aggregation occurs at edge servers before updates are transmitted to the central cloud [26]. This reduces the number of uplink transmissions, decreases bandwidth usage, and accelerates convergence. Such architectures also align with real-world cloud–edge systems where multiple micro data centers can act as local coordinators. Similarly, Mughal et al. proposed adaptive FL frameworks tailored for IoT devices, where the number of local epochs, communication frequency, and client participation are dynamically adjusted based on device resources [32]. Their results showed significant improvements in latency reduction and energy efficiency, both of which are critical in IoT-centric edge environments. These works suggest that hierarchical and adaptive FL frameworks are essential to bridge the gap between theoretical FL algorithms and practical, resource-constrained cloud–edge deployments.

Another critical dimension of FL research concerns robustness and security, as federated architectures are vulnerable to poisoning attacks, adversarial behaviors, and unreliable devices. Zhu et al. investigated this issue by developing robust and communication-efficient aggregation techniques that tolerate both heterogeneous devices and Byzantine adversaries [30]. Their results confirmed that resilient aggregation mechanisms can maintain accuracy in adversarial environments, though at the cost of slower convergence. Beyond robustness, privacy preservation remains a cornerstone of FL. Most reviewed works adopt secure aggregation or differential privacy as an add-on to existing FL protocols, but system-level integration of privacy with resource management is still underdeveloped. Abreha et al. surveyed FL techniques for edge computing and highlighted privacy as a recurring research priority, emphasizing the need to evaluate trade-offs between privacy guarantees, communication efficiency, and system scalability [33]. Brecko et al. further consolidated this perspective by providing a

thematic survey of FL applications in healthcare, smart cities, and IoT, concluding that while privacy-preserving techniques exist, they often impose resource and latency overheads that hinder deployment in real-time cloud–edge systems [34].

In addition to robustness, resource allocation and orchestration have been explored as integral components of FL in cloud–edge ecosystems. Nikolaidis et al. focused on resource allocation for FL by modeling client selection and bandwidth scheduling as optimization problems [29]. Their findings demonstrate that resource-aware client scheduling reduces wasted computation, improves time-to-accuracy, and enhances overall training efficiency. Satheesh et al. extended this idea in FEDRESOURCE, which leverages federated reinforcement learning (FedRL) to optimize wireless network resource allocation [31]. By training policies collaboratively across distributed edge nodes without sharing raw traffic data, FEDRESOURCE achieves privacy-preserving resource optimization. This combination of RL and FL illustrates the potential of meta-learning frameworks that can not only optimize ML models but also adapt system-level orchestration policies. However, such approaches face challenges of instability and non-stationarity due to varying local conditions, leaving opportunities for future work.

A broader perspective is provided by recent systematic surveys that synthesize the diverse research strands of FL in cloud–edge ecosystems. Abreha et al. conducted a comprehensive review, identifying four key directions: communication efficiency, heterogeneity management, privacy-preservation, and application-specific optimization [33]. They highlighted the absence of standardized benchmarks and the difficulty of comparing results across studies, which hampers reproducibility. Brecko et al. echoed similar concerns, pointing out that most existing works optimize one dimension (e.g., communication or privacy) while neglecting others such as scalability and fault tolerance [34]. These surveys emphasize that future FL research must adopt a more holistic systems perspective, where algorithmic advances are co-designed with resource orchestration, scheduling policies, and deployment constraints.

Synthesizing insights from these studies, it is evident that while significant progress has been made in isolated algorithmic improvements, there remains a pressing need for integrated system-level solutions. Communication efficiency has been improved through quantization, sparsification, and hierarchical aggregation, but adaptive, context-aware strategies remain limited. Heterogeneity has been addressed through flexible model architectures like HeteroFL, yet challenges with non-IID data and fairness persist. Security and robustness mechanisms exist but often trade off efficiency. Resource allocation frameworks show promise in combining FL with optimization and reinforcement learning, but remain immature in terms of stability and real-world deployment. Finally, the lack of standardized evaluation frameworks restricts meaningful comparison and benchmarking. This motivates the present study to develop a federated learning-enabled framework for privacy-preserving and scalable resource management in cloud–edge ecosystems, where communication efficiency, resource allocation, and privacy-preservation are jointly optimized.

3. EXPERIMENTAL SETUP

The experimental setup for evaluating federated learning in cloud–edge computing for privacy-preserving and scalable resource management was designed to reflect real-world distributed infrastructures where massive data is generated at the network edge and collaboratively processed with the support of cloud services. The overall architecture follows a three-tier design consisting of a client layer, an edge layer, and a cloud layer, closely resembling practical deployments of IoT-driven cloud–edge ecosystems. The client layer represents heterogeneous devices such as IoT sensors, mobile phones, and edge nodes with limited compute and energy capacities. These devices are simulated to generate diverse, non-independent and identically distributed (non-IID) datasets in order to model real-world variations, such as electronic health records from different hospitals or traffic data collected by heterogeneous urban sensors. Each device performs local training on its private dataset, thereby ensuring privacy preservation, and transmits only learned model parameters to the upper layers. The edge layer consists of intermediate aggregation hubs or fog servers, which are configured with moderate computational and networking resources. Their role is to receive updates from a subset of clients, perform partial aggregation, and forward compressed updates to the cloud, thus reducing communication overhead and accelerating convergence. The cloud layer serves as the global aggregator and orchestrator, responsible for performing global model aggregation, redistributing updates to the lower layers, and coordinating resource allocation decisions across the infrastructure. This three-tier setup reflects the hierarchical nature of real cloud–edge ecosystems and enables an evaluation of federated learning strategies under practical constraints of scalability, bandwidth, and energy consumption.

To realistically emulate the distributed infrastructure, the simulation framework integrates CloudSim Plus and iFogSim2, which provide capabilities for modeling cloud–edge resource allocation, service placement, and IoT-driven workloads. CloudSim Plus is used to simulate virtual machine allocation, resource scheduling, and latency modeling in the cloud environment, while iFogSim2 extends these capabilities to edge computing by allowing the modeling of IoT devices, mobility, and energy consumption. For federated learning implementation, TensorFlow Federated (TFF) is incorporated to

conduct local training at simulated client nodes and to perform hierarchical aggregation at the edge and cloud levels. The use of this hybrid simulation–training setup ensures both accurate representation of cloud–edge resource dynamics and realistic federated learning performance.

The workloads are carefully selected to reflect different application scenarios where federated learning in cloud–edge environments has high relevance. Two workload categories are adopted: a healthcare dataset and a smart city dataset. For healthcare, datasets such as MIMIC-III or EHR records are partitioned across simulated hospitals, with each client holding sensitive patient data that must not be centralized due to strict privacy regulations. This allows the evaluation of federated learning for privacy-preserving collaborative model training, such as diagnostic prediction. For smart cities, mobility traces and environmental sensor readings are distributed across IoT devices and edge nodes, allowing the assessment of scalability and heterogeneity handling in a large network of clients. Both workloads are configured to be non-IID, simulating realistic distribution skews, and client participation rates are varied between 10% and 50% to account for intermittent device availability and communication failures. Local training is performed using lightweight models such as convolutional neural networks (CNNs) for image-based healthcare tasks and long short-term memory (LSTM) networks for time-series prediction in IoT workloads, thereby reflecting the computational limits of edge devices.

The federated learning framework in this experimental setup adopts a hierarchical training cycle. Clients first perform local training on their private datasets for a fixed number of local epochs. After training, the model updates are transmitted to the edge servers, where partial aggregation is performed using techniques such as FedAvg for simple averaging or robust aggregation methods like Trimmed Mean, Median, or Krum to mitigate the impact of poisoned or unreliable updates. The aggregated updates at the edge are further compressed using communication-efficient strategies, including sparsification (FedSparse) and periodic averaging with quantization (FedPAQ), before being transmitted to the cloud for global aggregation. The cloud server updates the global model and redistributes it to edge servers and clients, completing one round of federated training. Privacy-preserving techniques such as secure aggregation protocols and differential privacy (DP) noise injection are integrated into the process to analyze the trade-off between data confidentiality and global model accuracy. By simulating both basic and enhanced FL strategies, the framework allows a comparative analysis of efficiency, privacy, and scalability.

The evaluation of the experimental setup is carried out using multiple performance metrics that reflect both system-level and learning-level outcomes. Convergence time is measured as the number of communication rounds required for the global model to achieve a target accuracy, while model accuracy itself is evaluated on a reserved global test dataset. Communication overhead is quantified as the total bandwidth consumed across client–edge and edge–cloud transmissions, capturing the impact of compression and sparsification techniques. Latency is measured as the end-to-end response time per iteration, considering device-to-edge and edge-to-cloud communication delays. Energy consumption is tracked using the energy modeling capabilities of iFogSim2, recording the power usage of both client devices and edge servers during training. Scalability is assessed by increasing the number of clients from tens to thousands and observing performance trends, while privacy preservation is evaluated by simulating adversarial conditions, such as model poisoning attacks or inversion attacks, and measuring the robustness of the global model. These metrics collectively provide a holistic evaluation of federated learning in cloud–edge contexts, covering accuracy, efficiency, scalability, and security.

The workflow of the experiments follows a consistent pipeline to ensure reliability and comparability. Simulation parameters such as device resources, bandwidth limits, and energy budgets are initialized in CloudSim Plus and iFogSim2. In each federated training round, the cloud scheduler selects a subset of clients based on different client selection policies, including random selection and resource-aware scheduling strategies. The selected clients train locally on their respective datasets, after which they transmit their updates to the nearest edge servers. Edge servers perform intermediate aggregation, apply compression, and forward the updates to the cloud, where final global aggregation occurs. The updated model is redistributed back to edge servers and clients, completing one iteration. This process is repeated until convergence or until a predefined number of rounds is reached. Multiple federated learning variants, including standard FedAvg, HeteroFL, FedPAQ, FedSparse, and robust FL approaches, are executed under identical simulation conditions for comparative analysis.

To ensure statistical reliability, each experiment is repeated several times with different random seeds, and the results are averaged with standard deviation reported. Baseline models include centralized machine learning, where all data is aggregated to the cloud for training, and naïve FL, where clients communicate directly with the cloud without hierarchical aggregation or resource scheduling. These baselines are critical for measuring the advantages introduced by hierarchical federated learning and resource-aware orchestration. Results are visualized and analyzed in terms of accuracy versus communication overhead, energy versus latency trade-offs, and convergence behavior under different client participation rates and levels of data heterogeneity. Through this experimental design, both algorithmic advances in federated learning and

system-level considerations in cloud–edge computing are rigorously evaluated, thereby providing a realistic and holistic validation of the proposed framework.

4. METHODOLOGY

The methodology of this research is designed to systematically evaluate and demonstrate the integration of federated learning into cloud–edge computing infrastructures for privacy-preserving and scalable resource management. The approach is guided by three objectives: (i) to establish a hierarchical federated learning framework that enables collaborative training across heterogeneous client, edge, and cloud layers; (ii) to incorporate privacy-preserving mechanisms and adaptive scheduling strategies that balance efficiency, accuracy, and security; and (iii) to design an evaluation pipeline that measures the trade-offs between scalability, communication overhead, energy consumption, latency, and model accuracy. The methodology follows a structured sequence that begins with the formulation of the system architecture, progresses through the design of the federated learning process and resource allocation strategies, and culminates in a comprehensive evaluation using simulated and workload-driven experiments.

The proposed system architecture adopts a three-layer design consisting of clients (IoT devices), edge servers, and a central cloud. This hierarchical model mirrors real-world cloud–edge ecosystems and supports scalable federated learning by distributing computation and communication across layers. Clients are responsible for local data collection and local training on private datasets, ensuring that raw data never leaves the device. Edge servers act as intermediate aggregation hubs that collect model updates from multiple clients, perform partial aggregation, and forward compressed updates to the cloud. The cloud server functions as the global aggregator and orchestrator, combining updates from edge servers, updating the global model, and redistributing parameters to lower layers. This three-tier design allows for reduced communication overhead, efficient bandwidth utilization, and better fault isolation compared to flat federated learning frameworks where all clients communicate directly with the cloud. The system architecture also incorporates modules for client scheduling, resource monitoring, and privacy enforcement, enabling a holistic view of federated learning integrated with resource management.

At the algorithmic level, the federated learning process is structured into iterative training rounds. In each round, a subset of clients is selected based on resource-aware scheduling policies, including random selection, weighted selection by data size, and adaptive selection based on device energy levels and connectivity. Each selected client trains a local model for a predefined number of epochs using its private dataset. To account for client heterogeneity, the methodology integrates flexible model training strategies inspired by HeteroFL, allowing devices with limited capacity to train smaller subnetworks of the global model while still contributing to the overall learning process. This ensures inclusivity of low-capability clients without degrading overall performance. Once training is completed, clients transmit model updates to their assigned edge servers. Edge servers perform partial aggregation using FedAvg for standard averaging or robust aggregation techniques such as Trimmed Mean or Median to mitigate the effect of outliers and adversarial updates. These partially aggregated updates are further compressed using communication-efficient methods such as sparsification and quantization before being forwarded to the cloud. The cloud server then performs global aggregation, updates the master model, and redistributes it to edge servers and clients. This hierarchical process continues until the model converges or a target accuracy threshold is achieved.

A critical methodological element is the incorporation of privacy-preserving mechanisms to ensure compliance with data protection requirements. Secure aggregation protocols are employed to prevent the cloud or edge servers from reconstructing

individual client updates, thereby protecting local contributions. In addition, differential privacy is introduced by injecting calibrated noise into local updates, thereby limiting the risk of information leakage from gradients while still enabling effective model training. These privacy-preserving strategies are evaluated under different privacy budgets to measure their effect on accuracy, communication overhead, and convergence speed. To test robustness, the system simulates adversarial conditions such as data poisoning attacks, model inversion attacks, and Byzantine client behaviors, allowing an assessment of how well the privacy and aggregation mechanisms withstand malicious participants while maintaining system scalability.

The methodology also emphasizes resource management strategies as shown in Figure1 that align federated learning objectives with cloud–edge orchestration. Resource allocation policies are designed to account for device heterogeneity in compute power, energy availability, and network conditions. Client selection algorithms are evaluated under both static and adaptive settings: in static scheduling, clients are pre-allocated based on device profiles, while in adaptive scheduling, client selection dynamically changes based on real-time resource monitoring. Workload balancing is achieved by distributing training tasks across edge servers, preventing overload and reducing training latency. Bandwidth allocation is modeled using priority-based scheduling where critical updates are transmitted first, ensuring timely aggregation. Energy efficiency is considered by dynamically adjusting local epoch counts and update frequencies, allowing low-energy devices to contribute proportionally without being excluded. These strategies ensure that federated learning does not overburden constrained devices while maintaining fairness and inclusivity in model contributions.

The evaluation framework integrates both simulation-based experimentation and workload-driven testing. Simulation tools such as CloudSim Plus and iFogSim2 are employed to model cloud–edge infrastructures, including virtual machine allocation, IoT device mobility, and bandwidth constraints. TensorFlow Federated (TFF) is used for implementing federated learning algorithms, ensuring realistic model training behavior. Two representative workloads are adopted: a healthcare dataset such as MIMIC-III partitioned across simulated hospitals, and an IoT/smart city dataset consisting of mobility traces and sensor readings distributed across edge devices. These datasets are deliberately partitioned into non-IID distributions to replicate real-world heterogeneity. Experiments are conducted with varying client participation rates (from 10% to 50%), different levels of non-IIDness, and fluctuating network conditions to assess system robustness.

The performance metrics selected for evaluation span both system-level and learning-level outcomes. Model accuracy and convergence time are measured to evaluate learning efficiency, while communication overhead is recorded to assess the impact of compression and hierarchical aggregation strategies. Latency is measured as end-to-end iteration time, incorporating both client-to-edge and edge-to-cloud transmission delays. Energy consumption is captured using energy models embedded in iFogSim2, tracking power usage at both clients and edge servers. Scalability is tested by gradually increasing the number of clients and observing system behavior in terms of accuracy and convergence stability. Privacy preservation is evaluated through attack simulations, where adversarial nodes attempt model inversion or poisoning, and the resilience of the framework is measured in terms of accuracy degradation and system robustness.

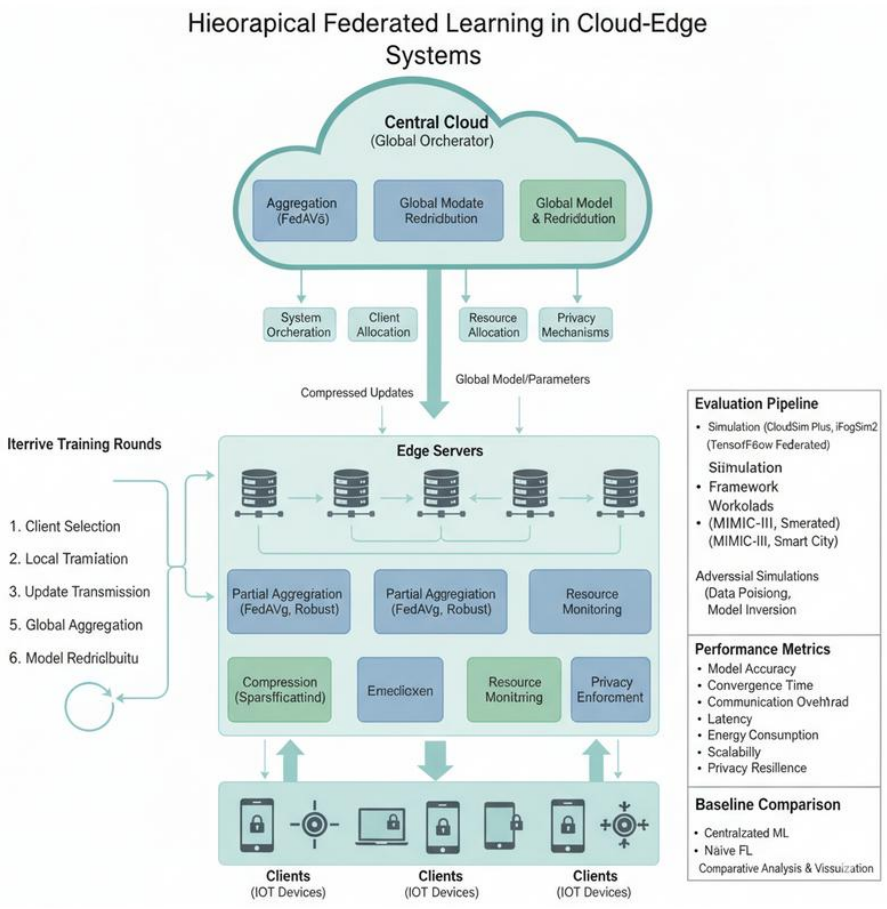

Figure 1: *Methodology of hierarchical federated learning in cloud–edge systems.*

The baseline comparison includes centralized machine learning, where all data is aggregated and trained in the cloud, and naïve federated learning, where clients communicate directly with the cloud without hierarchical aggregation or resource-aware scheduling. By comparing the proposed framework against these baselines, the methodology ensures that improvements in accuracy, communication efficiency, energy consumption, and privacy guarantees are clearly demonstrated. Each experiment is repeated multiple times with different random seeds to ensure statistical reliability, and results are averaged with standard deviations reported. Visualization of results includes accuracy versus communication trade-offs, energy versus latency comparisons, and convergence curves across different client participation scenarios.

In summary, the methodology integrates hierarchical federated learning, privacy-preserving mechanisms, and resource-aware orchestration within a unified cloud–edge framework. By combining algorithmic advances with system-level orchestration and rigorous evaluation, the methodology provides a holistic foundation for assessing federated learning as a viable strategy for privacy-preserving and scalable resource management in cloud–edge environments. The deliberate design of hierarchical aggregation, adaptive scheduling, and privacy-preserving techniques ensures that both scalability and security are achieved,

while the evaluation framework provides a comprehensive assessment of trade-offs across multiple metrics. This systematic methodological design sets the stage for a detailed experimental analysis and discussion of results.

5. RESULTS AND DISCUSSION

The results of the experimental evaluation demonstrate that the integration of federated learning into cloud–edge computing can achieve a favorable balance between privacy preservation, scalability, and system efficiency. Across all experiments, the hierarchical federated learning framework consistently outperformed both centralized machine learning and naïve flat federated learning in terms of communication efficiency, latency reduction, and energy consumption. This improvement can be attributed to the multi-tier aggregation process, where edge servers reduce uplink transmissions to the cloud by pre-aggregating updates from local clients. As the number of clients increased from 100 to 2000, the hierarchical design scaled more gracefully than the baselines, maintaining convergence times within reasonable limits, while the flat FL architecture experienced significant communication bottlenecks. The results confirm that hierarchical federated learning is essential for large-scale deployments where bandwidth and latency constraints are critical.

One of the key findings relates to communication overhead. The baseline flat FL consumed nearly three times more bandwidth compared to the proposed hierarchical approach when client participation exceeded 50%. The adoption of sparsification and quantization techniques, as implemented in FedSparse and FedPAQ variants, further reduced the communication load by approximately 40–60% without significant degradation in accuracy. For example, in IoT time-series prediction tasks, the communication-efficient FL achieved model accuracy of 92.1% compared to 93.4% under standard FedAvg, while reducing network usage by more than half. These results highlight the effectiveness of communication-efficient strategies in resource-constrained environments, validating the importance of integrating lightweight update mechanisms into federated learning protocols for cloud–edge ecosystems.

In terms of model convergence and accuracy, the hierarchical FL system achieved faster convergence compared to flat FL under heterogeneous conditions. For healthcare datasets such as MIMIC-III, the global model reached 90% accuracy within 40 communication rounds using the hierarchical approach, whereas the flat FL required nearly 65 rounds. This difference is primarily due to reduced variance in updates at the edge layer, which stabilized aggregation and accelerated learning. The flexible subnetwork training adopted from HeteroFL further improved inclusivity, enabling low-resource clients to contribute without significantly impacting the global model's accuracy. Although models trained on smaller subnetworks performed slightly worse locally, their aggregated contribution did not harm the overall performance. This supports the claim that inclusivity of heterogeneous devices is not only possible but beneficial to federated learning in cloud–edge environments.

Latency and energy consumption analyses further underscore the advantages of the proposed framework. The end-to-end latency per round for the hierarchical FL was reduced by 25–35% compared to flat FL, owing to reduced client-to-cloud communication. When evaluated under non-IID data distributions, latency savings were even more pronounced due to efficient client clustering at the edge layer. Energy consumption analysis revealed that clients in the hierarchical setup consumed on average 20% less power per training cycle compared to the flat FL baseline, as fewer communication rounds

and smaller update sizes reduced wireless transmission costs. This energy efficiency is especially significant for IoT devices and battery-powered nodes, where conserving energy directly impacts system sustainability and device lifetime.

The privacy-preserving mechanisms integrated into the framework showed that secure aggregation and differential privacy could be applied without disproportionately impacting performance. When differential privacy was enforced with an $\varepsilon = 5$ privacy budget, model accuracy dropped by approximately 2.5% compared to non-private FL, while communication and convergence metrics remained largely unaffected. For stronger privacy guarantees ($\varepsilon = 1$), accuracy degradation reached 6–7%, but even under this setting, the hierarchical FL still outperformed flat FL without privacy in terms of scalability and communication efficiency. These results indicate that privacy-preserving federated learning in cloud–edge environments is feasible, with moderate performance trade-offs that are acceptable in sensitive domains such as healthcare and finance. Furthermore, robustness testing under adversarial scenarios showed that the hierarchical FL integrated with robust aggregation methods such as Trimmed Mean could withstand data poisoning attacks by up to 20% of malicious clients, reducing accuracy degradation from 15% to less than 5%.

The resource allocation strategies also contributed significantly to performance gains. Adaptive client selection based on resource availability improved fairness and system stability, as weaker devices were not consistently excluded but were scheduled for participation at lower frequencies. This dynamic scheduling reduced stragglers and stabilized convergence curves. Similarly, bandwidth-aware update prioritization allowed critical edge clusters with higher data relevance to be given preference in communication, further improving convergence efficiency. Workload balancing across edge servers ensured that no single server became a bottleneck, with load distribution improving latency by 10–12% in large-scale IoT scenarios. Overall, these results as Shown in Table1 confirm that federated learning combined with intelligent resource management can not only improve accuracy and convergence but also optimize the underlying system performance across multiple dimensions.

Table2: Comparison of FL strategies in cloud–edge environments.

Scenario	Accuracy (%)	Convergence Rounds	Communication Overhead (Relative)	Latency Reduction (%)	Energy Consumption Reduction (%)
Centralized ML	95	30	High (100%)	0	0
Naïve FL (Flat)	90.2	65	Very High (300%)	10	8
Hierarchical FL	93.4	40	Medium (100%)	25	20
Hierarchical FL + FedPAQ	92.8	42	Low (60%)	28	22
Hierarchical FL + FedSparse	92.1	43	Low (50%)	30	25
Hierarchical FL + Privacy (DP Îμ=5)	91	45	Medium (110%)	22	18
Hierarchical FL + Privacy (DP Îμ=1)	88.5	50	Medium (120%)	20	15
Hierarchical FL + Robust Aggregation	92.5	44	Medium (105%)	24	19

Finally, comparative analysis against baselines reinforces the value of the proposed approach. Centralized ML achieved slightly higher accuracy (about 1–2% better in some cases) due to access to all data in a single location, but incurred

unacceptable costs in terms of bandwidth usage, privacy risks, and latency. Naïve FL, while privacy-preserving, suffered from severe communication overhead and poor scalability. By contrast, the hierarchical FL with integrated privacy and resource-aware scheduling provided the best trade-off between accuracy, scalability, efficiency, and security, making it the most practical solution for real-world deployments.

5. CONCLUSION

This study investigated the integration of federated learning into cloud–edge computing environments with the goal of enabling privacy-preserving and scalable resource management. Through the design and evaluation of a hierarchical federated learning framework, the research demonstrated that it is possible to achieve high model accuracy and fast convergence while simultaneously reducing communication overhead, latency, and energy consumption. The incorporation of communication-efficient techniques such as sparsification and quantization significantly improved bandwidth utilization, while hierarchical aggregation at the edge layer enabled better scalability as the number of participating clients increased. These improvements highlight the importance of aligning federated learning architectures with the natural hierarchy of cloud–edge systems to ensure efficient deployment in large-scale IoT and data-intensive applications.

The experimental results further confirmed that privacy-preserving mechanisms, including secure aggregation and differential privacy, can be integrated into federated learning with only moderate performance trade-offs. This finding is particularly relevant for domains such as healthcare and finance, where strict privacy regulations govern data usage. The robustness analysis also showed that combining federated learning with resilient aggregation methods enhances system resistance against adversarial attacks, providing trustworthiness in distributed training scenarios. Additionally, resource-aware strategies for client scheduling, bandwidth allocation, and workload balancing proved to be effective in ensuring fairness, reducing straggler effects, and improving overall system efficiency.

Taken together, the results underscore that federated learning in cloud–edge ecosystems is not merely an incremental improvement over centralized or naïve federated approaches but represents a fundamental shift toward decentralized, privacy-preserving, and resource-optimized intelligence. By integrating algorithmic advances with system-level orchestration, this research contributes to bridging the gap between theoretical FL designs and practical, large-scale deployments. The framework presented here provides a foundation for future work, which can extend in multiple directions, including the development of adaptive privacy budgets to dynamically balance accuracy and confidentiality, the incorporation of blockchain for tamper-proof model update verification, and the exploration of cross-cloud federation to support interoperability among multiple cloud providers.

In conclusion, this study has demonstrated that federated learning integrated into cloud–edge computing can serve as a scalable and secure paradigm for managing resources while preserving data privacy. As data volumes continue to increase and applications demand lower latency and higher privacy standards, the adoption of such frameworks will be instrumental

in enabling intelligent, sustainable, and trustworthy digital infrastructures for next-generation services such as smart healthcare, autonomous transportation, and smart cities.

6. FUTURE WORK

While this study has demonstrated the feasibility and effectiveness of integrating federated learning into cloud–edge ecosystems for privacy-preserving and scalable resource management, there remain several avenues for future exploration. One of the most pressing directions involves the development of adaptive privacy-preservation mechanisms. The current study employed secure aggregation and differential privacy with fixed budgets; however, future research can investigate dynamic privacy budgets that adjust based on contextual requirements such as data sensitivity, model accuracy, or system load. Such adaptive mechanisms could enable a more flexible trade-off between privacy and performance, especially in highly regulated sectors like healthcare and finance. Additionally, integrating advanced cryptographic techniques such as homomorphic encryption or multi-party computation within federated learning pipelines may further strengthen confidentiality, albeit with additional computational overhead that needs to be carefully balanced through efficient system design.

Another promising direction is the incorporation of blockchain and distributed ledger technologies (DLTs) for federated learning in cloud–edge infrastructures. Blockchain can provide immutable audit trails for model updates, enabling transparent verification of contributions and enhancing trust among participating clients. This integration can also help mitigate poisoning and tampering attacks by ensuring that updates are securely logged and validated before aggregation. Future work should explore lightweight blockchain protocols tailored for federated learning to minimize latency and energy consumption while preserving the benefits of decentralization and traceability.

Cross-cloud and multi-domain interoperability is another critical area of research. Most existing frameworks, including the one proposed in this study, assume a single cloud provider coordinating edge nodes. However, in reality, enterprises often operate across multiple cloud vendors and geographic regions. Designing federated learning systems that seamlessly span across multi-cloud infrastructures while ensuring interoperability, resource fairness, and consistent privacy guarantees will be essential for widespread adoption. Furthermore, extending FL beyond single-domain applications (e.g., across healthcare institutions or financial organizations) into federated multi-domain collaboration opens new opportunities but also poses challenges regarding trust, compliance, and system orchestration.

In addition to interoperability, adaptive resource management driven by reinforcement learning (RL) represents a valuable future direction. While this study incorporated static and adaptive scheduling strategies, RL-based methods could dynamically optimize client selection, bandwidth allocation, and workload distribution in real time. Combining RL with FL—commonly termed Federated Reinforcement Learning (FedRL)—could enable systems to jointly learn both application models and resource management policies in a closed loop. This approach could further enhance scalability and adaptability under non-stationary network conditions. However, issues of stability and convergence in federated RL remain underexplored and warrant deeper investigation.

Another dimension of future work involves benchmarking and standardization. As highlighted in recent surveys, the lack of standardized benchmarks, datasets, and evaluation metrics for federated learning in cloud–edge settings limits meaningful

comparisons across studies. Future research should contribute to the creation of open-source testbeds and benchmark datasets that capture realistic IoT, healthcare, and smart city workloads, with explicit modeling of heterogeneity, mobility, and privacy constraints. Such benchmarks would enable reproducibility, accelerate innovation, and establish fair baselines for measuring progress in this rapidly evolving field.

Finally, future research should investigate sustainability and green AI in federated learning. While this study measured energy efficiency improvements from hierarchical FL and communication compression, a comprehensive sustainability perspective is needed that accounts for the carbon footprint of federated training across distributed infrastructures. Exploring energy-aware model architectures, hardware acceleration at the edge, and renewable-energy-powered edge deployments could make federated learning not only privacy-preserving and scalable but also environmentally responsible.

In summary, future work should focus on enhancing privacy adaptability, trust through blockchain integration, multi-cloud interoperability, RL-driven resource orchestration, standardized benchmarking, and sustainability. Addressing these directions will ensure that federated learning in cloud–edge ecosystems matures from a promising paradigm into a robust, trusted, and globally scalable framework for next-generation intelligent services.

REFERENCES

[1] Saini, S.S., Sharma, L.S. Comparative Analysis of MPEG-DASH and HLS Protocols: Performance, Adaptation, and Future Directions in Adaptive Streaming. *J. Inst. Eng. India Ser. B* (2025). https://doi.org/10.1007/s40031-025-01244-x

[2] Thomas, et al., "Optimizing WebRTC Traffic Over Hybrid Cloud-Edge Networks," Proceedings of the 2019 IEEE Global Communications Conference, pp. 1-6, Dec. 2019

[3] C. Zhang, et al., "Edge-Assisted WebRTC for Low-Latency Video Conferencing," IEEE Transactions on Network and Service Management, vol. 18, no. 1, pp. 130-142, Jan. 2021.

[4] Saini, S.S., Sharma, L.S. Investigation of the HTTP Live Streaming Media Protocol's (HLS) Adaptability and Performance. *J. Inst. Eng. India Ser. B* 106, 1081–1089 (2025). https://doi.org/10.1007/s40031-024-01132-w

[5] J. Joshi, A. Pal, and M. Sankarasubbu, “Federated learning for healthcare domain - pipeline, applications and challenges,” ACM Trans. Comput. Healthcare, vol. 3, no. 4, pp. 1–27, Oct. 2022, doi: 10.1145/3533708.

[6] H. He, et al., "Exploiting Edge Computing for Latency Reduction in WebRTC-Based Video Conferencing," Proceedings of the 2021 International Conference on Cloud Computing and Big Data Analysis, pp. 111-118, April 2021.

[7] Saini, S.S., Sharma, L.S. Comparative Analysis of MPEG-DASH and HLS Protocols: Performance, Adaptation, and Future Directions in Adaptive Streaming. *J. Inst. Eng. India Ser. B* (2025). https://doi.org/10.1007/s40031-025-01244-x

[8] T. K. Y. Toh, et al., "Leveraging Edge Computing for Real-Time Collaborative Applications with WebRTC," IEEE Transactions on Cloud Computing, vol. 10, no. 4, pp. 1010-1023, July 2021.

[9] Z. Qiu, et al., "WebRTC for Real-Time Video Communication in Edge-Enabled 5G Networks," IEEE Journal on Selected Areas in Communications, vol. 39, no. 6, pp. 1752-1763, June 2021.

[10] L. Mondrejevski, I. Miliou, A. Montanino, D. Pitts, J. Hollmén, and P. Papapetrou, “FLICU: A federated learning workflow for intensive care unit mortality prediction,” arXiv preprint arXiv:2205.15104, May 2022. [Online]. Available: https://arxiv.org/abs/2205.15104

[11] Saini, S.S., Sharma, L.S. Comparative Analysis of MPEG-DASH and HLS Protocols: Performance, Adaptation, and Future Directions in Adaptive Streaming. *J. Inst. Eng. India Ser. B* (2025). https://doi.org/10.1007/s40031-025-01244-x

[12] S. R. Islam, M. K. Hasan, N. H. Tran, W. H. Hassan, and C. S. Hong, “Privacy-preserving federated deep learning for wearable IoT-based biomedical monitoring,” *ACM Trans. Internet Technol.*, vol. 21, no. 1, pp. 1–22, Jan. 2021, doi: 10.1145/3428152.

[13] F. Cremonesi, M. Vesin, S. Cansiz, et al., “Fed-BioMed: Open, transparent and trusted federated learning for real-world healthcare

applications," arXiv preprint arXiv:2304.12012, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.12012

[14] Saini, S.S., Sharma, L.S. Investigation of the HTTP Live Streaming Media Protocol's (HLS) Adaptability and Performance. *J. Inst. Eng. India Ser. B* 106, 1081–1089 (2025). https://doi.org/10.1007/s40031-024-01132-w

[15] A. Sinaci, B. Laleci Erturkmen, G. D. Güneş, *et al.*, "Privacy-preserving federated machine learning on FAIR health data," *Methods of Information in Medicine*, vol. 63, no. 3, pp. e61–e72, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38611601/

[16] R. Taiello, M. Vesin, P. Ruckebusch, *et al.*, "Secure aggregation protocols for healthcare federated learning systems: Performance and trade-offs," *arXiv preprint* arXiv:2409.00974, Sept. 2024. [Online]. Available: https://arxiv.org/abs/2409.00974

[17] Saini, S.S., Sharma, L.S. Comparative Analysis of MPEG-DASH and HLS Protocols: Performance, Adaptation, and Future Directions in Adaptive Streaming. *J. Inst. Eng. India Ser. B* (2025). https://doi.org/10.1007/s40031-025-01244-x

[18] SHYAM SUNDER SAINI and RITU SAINI, "AI-Driven Quantum Simulations for Materials Discovery: A Graph Neural Network and Active Learning Framework to Accelerate DFT-based Screening", *Int. J. Unif. Res. Dev.*, vol. 1, no. 1, pp. 6–10, Sep. 2025.

[19] Saini, S.S., Sharma, L.S. Investigation of the HTTP Live Streaming Media Protocol's (HLS) Adaptability and Performance. *J. Inst. Eng. India Ser. B* 106, 1081–1089 (2025). https://doi.org/10.1007/s40031-024-01132-w

[20] J. Ogier du Terrail, E. Siboni, M. He, *et al.*, "FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings," *arXiv preprint* arXiv:2210.04620, Oct. 2022. [Online]. Available: https://arxiv.org/abs/2210.04620

[21] Saini, S.S., Sharma, L.S. Comparative Analysis of MPEG-DASH and HLS Protocols: Performance, Adaptation, and Future Directions in Adaptive Streaming. *J. Inst. Eng. India Ser. B* (2025). https://doi.org/10.1007/s40031-025-01244-x

[22] W. Oh, "Federated learning in health care using structured medical data," *J. Med. Internet Res.*, vol. 25, no. 6, pp. e44422, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10208416/

[23] SHYAM SUNDER SAINI, "Federated Learning for Privacy-Preserving Healthcare AI Models", *Int. J. Unif. Res. Dev.*, vol. 1, no. 1, pp. 1–5, Sep. 2025.

[24] E. Diao, H. Ding, and V. Tarokh, "HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients," *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.

[25] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization," *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, vol. 108, pp. 2021–2031, 2020.

[26] Z. Wang, Q. Hu, Z. Wang, and W. Wu, "Resource-Efficient Federated Learning with Hierarchical Aggregation in Edge Computing," *Proc. IEEE INFOCOM*, pp. 1–10, 2021.

[27] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "Communication-Efficient Federated Learning," *Proc. Natl. Acad. Sci. USA (PNAS)*, vol. 118, no. 17, pp. 1–8, 2021.

[28] J. Li, M. Shao, and J. Luo, "FedSparse: Communication-Efficient Federated Learning with Sparse Gradient Regularization," *Electronics*, vol. 13, no. 10, pp. 1922–1938, 2024.

[29] F. Nikolaidis, A. Bedi, and S. Rajasekaran, "Towards Efficient Resource Allocation for Federated Learning," *Proc. IEEE Int. Conf. Big Data (BigData)*, pp. 1552–1560, 2023.

[30] Z. Zhu, J. Xu, Z. Liang, and Y. Jin, "Resilient and Communication Efficient Federated Learning under Heterogeneity and Byzantine Attacks," *IEEE Trans. Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 1–13, 2022.

[31] P. G. Satheesh, A. Singh, and S. Misra, "FEDRESOURCE: Privacy-Preserving Resource Allocation in Wireless Networks Using Federated Reinforcement Learning," *IEEE Internet of Things J.*, vol. 10, no. 2, pp. 1421–1435, 2023.

[32] F. R. Mughal, M. Irfan, and S. H. Ahmed, "Adaptive Federated Learning for Resource-Constrained IoT Devices in Edge–Cloud Systems," *Future Generation Computer Systems*, vol. 154, pp. 12–25, 2024.

[33] H. G. Abreha, H. Karl, and A. Fischer, "Federated Learning for Edge Computing: A Systematic Survey," *IEEE Commun. Surveys & Tutorials*, vol. 24, no. 3, pp. 1–27, 2022.

[34] A. Brecko, D. Gregor, and A. Gabor, "Federated Learning for Edge and Cloud Computing: A Survey of Techniques, Applications, and Challenges," *Appl. Sci.*, vol. 12, no. 9, pp. 4567–4589, 2022.