Received: 27 May 2025 | Revised: 8 August 2025 | Accepted: 17 October 2025 | Published online: 15 November 2025

REVIEW

Artificial Intelligence and Applications
2025, Vol. 00(00) 1–5
DOI: 10.47852/bonviewAIA52026275

BON VIEW PUBLISHING

Communication-Efficient Federated Learning: A Systematic Review of Model Compression and Aggregation Techniques

Femi Temitope Johnson[1,*], Elugbadebo Oladapo[2], Olukumoro Olugbenga[3], Alomaja Victor[3], and Akande Adenike[2]

[1] *Department of Computer Science, Federal University of Agriculture, Nigeria*

[2] *Computer Science Department, Federal College of Education, Nigeria*

[3] *Department of Computer Science, Yaba College of Technology, Nigeria*

Abstract: Federated Learning (FL) has emerged as a transformative paradigm for decentralized machine learning, enabling collaborative model training across distributed devices while preserving data privacy. However, high communication overhead during model updates remains a major obstacle to real-world deployment. This paper presents a systematic literature review aimed at identifying and evaluating communication-efficient techniques in FL, with a focus on model compression and aggregation methods. Using a PRISMA-based framework, we reviewed 65 peer-reviewed studies published between 2018 and 2024 from IEEE Xplore, ACM Digital Library, and Scopus. The analysis reveals key contrasts between research contexts: 92% of industry-led studies emphasize reproducibility and deployment readiness, compared to 61% of academic publications. Among the most effective methods, quantization techniques offer 40–70% bandwidth savings, while advanced aggregation strategies like FedProx and SCAFFOLD enhance performance under data heterogeneity. Despite these advances, fundamental trade-offs remain between compression ratio, convergence speed, and model accuracy. Additionally, reproducibility is hindered by inconsistent benchmarks and limited open-source tools. Gaps in energy-efficient protocols, sustainability metrics, and cross-domain adaptability also constrain deployment in edge and IoT environments. This review provides a foundation for developing scalable, reproducible, efficient and communication-aware FL systems for practical applications.

Keywords: federated learning, communication efficiency, model compression, gradient aggregation, edge computing

1. Introduction

The rapid increase in the data generated by edge devices like Smartphone, Internet of Things (IoT) sensors, and wearables has revolutionized the training and implementation of machine learning models. Traditional centralized learning methods necessitate the transmission of raw data [1] to a central server, creating considerable challenges regarding privacy, security, and bandwidth [2, 3]. In response to the challenges, Federated Learning (FL) has surfaced as a decentralized option that allows collaborative model training across various devices while keeping raw data private [4, 5]. By maintaining sensitive information locally, FL tackles major issues associated with centralized data gathering, safeguarding user privacy [6] and adhering to data protection laws like GDPR [7, 8]. This framework entails several clients collaboratively training a shared global model with their local datasets, periodically transmitting model updates to a central server for aggregation, frequently utilizing methods like Federated Averaging (FedAvg) [9].

In addition to privacy, FL seeks to decrease the use of network bandwidth and latency [10, 11], especially in environments with limited resources and connectivity. Its uses are growing swiftly in healthcare [12–15], finance [16], and mobile device customization [17, 18], where sensitivity and diversity of data are significant issues. The rise of edge computing and IoT devices has enhanced the need for a decentralized machine learning models that maintain a balance between privacy and performance [19, 20].

FL has consequently surfaced as a novel framework, allowing cooperative training on edge devices while maintaining raw data in its local environment [21]. Despite its theoretical advantages, FL faces significant practical challenges [22], including high communication overhead due to iterative model updates between distributed devices and central servers [23]. Recent studies indicate that communication costs account for 70–90% of total FL system latency, representing a major bottleneck in real-world deployment [24]. Additionally, FL introduces challenges related to data heterogeneity [25], system scalability [26], and security threats such as poisoning and inference attacks. These unresolved issues continue to drive research aimed at improving FL's robustness, scalability, and privacy preservation [27, 28].

Modern methods aimed at improving FL communication efficiency concentrate on two complementary tactics: (1) model compression methods such as quantization [29], pruning [30, 31], and gradient sparsification [32, 33]; and (2) enhanced aggregation algorithms like FedProx [22] and SCAFFOLD [34]. Although these approaches demonstrate potential, there is a deficiency in a thorough synthesis of their overall effectiveness, practical constraints, and best applications. For instance, although 1-bit quantization can decrease bandwidth needs by 60–70%, it frequently falls short on non-IID data [35]. In the same way, sophisticated aggregation methods suggested by Wang et al. [36] enhance convergence in diverse FL settings, though they might lead to increased computational complexity.

This systematic review directly tackles three significant deficiencies in existing FL studies. Initially, there is no extensive framework for assessing how compression methods relate to different aggregation algorithms in FL systems [37].

***Corresponding author:** Femi Temitope Johnson, Department of Computer Science, Federal University of Agriculture, Nigeria. femijohnson@funaab.edu.ng

Secondly, current research has not sufficiently defined the practical compromises among communication efficiency, model accuracy, and convergence rates in various FL contexts [38]. Third, a significant lack of standardized benchmarks persists for objectively evaluating communication efficiency enhancements across various FL implementations [39, 40]. By meticulously examining 65 peer-reviewed studies released from 2018 to 2024, this review offers four significant contributions to the progress of communication-efficient FL research.

1) Development of a novel taxonomy classifying FL optimization techniques based on compression-aggregation co-design principles.
2) Quantitative meta-analysis of bandwidth-accuracy trade-offs across 12 major FL algorithms.
3) Identification of previously understudied challenges in dynamic network environments and highly heterogeneous device ecosystems.
4) Evidence-based guidelines for technique selection based on specific deployment constraints and performance requirements.

Also based on the listed contributions the following objectives and questions have been formulated to strengthen the focus and coherence of this systematic review.

1.1. Research objective

This review aims to identify, categorize, and critically analyze recent communication-efficient techniques in FL, with a focus on industrial versus academic approaches, and the translation of theoretical models into real-world deployments

1.2. Research questions

RQ1: What are the dominant strategies used to reduce communication overhead in FL since 2020?
RQ2: How do academic and industrial implementations of communication-efficient FL differ in terms of design, evaluation, and reproducibility?
RQ3: What gaps and challenges exist in current communication-efficient FL approaches that hinder scalable, sustainable deployment?

The remainder of this paper is organized as follows: Section 2 presents a literature review while Section 3 details our PRISMA-compliant methodology and study selection process. Section 4 presents our systematic analysis of compression and aggregation techniques, while Section 5 discusses emerging challenges and future research directions. Finally, Section 6 concludes with practical recommendations for researchers and practitioners implementing communication-efficient FL systems.

2. Review of Related Works

Prior to 2020, several foundational studies played a pivotal role in shaping communication-efficient strategies. Various techniques in distributed optimization were also explored to enhance communication efficiency among devices including Daume's distribution protocol [41]. This method utilized a multiplicative weight update, facilitated effective classification and optimization of data shared among nodes in a distributed network. While their protocol was simple, efficient, and well-suited for convex problems, it faced limitations in communication, as costs increased with dimensionality. Despite reducing the number of agents and their dimensionality compared to naive methods where full datasets and gradients were transferred without privacy preservation their approach struggled with scalability in high-dimensional settings.

As communication efficiency became increasingly critical in real-world systems, researchers introduced the Communication-Efficient Distributed Dual Coordinate Ascent (CoCoA) framework [42], which facilitated asynchronous updates on a central server. This framework generated a global model update for each communication device, proving adaptable to a wide range of optimization problems. Their results demonstrated a reduction in communication issues by at least 10% while maintaining improved accuracy. However, the framework's effectiveness heavily relied on the design of local solvers, highlighting a key dependency.

The exploration of the impact of quantization on device communication performed by Yuan et al. [43], demonstrated and achieved a slow but formal convergence even with low-precision updates in multi-agent systems. A major advantage of their technique was its ability to control quantization noise in both synchronous and asynchronous environments, ensuring stability across different communication settings. A novel triple-factor balancing algorithm was deployed in the study from Agarwal et al. [44], to optimize communication efficiency, user privacy, and model accuracy.

Their method utilized single-bit gradient quantization with a privacy mechanism, which also eliminated biases to ensure stable convergence. Notably, the technique significantly reduced communication data size without substantial performance loss in tested models, striking a balance between efficiency and accuracy.

In 2014, Li et al. [45] successfully integrated machine learning through a parameter architecture that supported timely updates and synchronization of communication data. This technique proved highly effective in production environments, enabling real-time learning, resource management, and robust communication filtering. It achieved significant bandwidth savings on the client side and an approximately 40× compression rate on the server without degrading model performance. However, it suffered from consistency issues due to delayed updates, presenting a notable drawback.

To address decentralized learning challenges, an experiment was performed [46], with a decentralized stochastic algorithm designed to solve root-finding monotone operator problems that hindered sparse communication among nodes. Their findings revealed that geometric convergence was achievable through convex minimization and AUC maximization. Remarkably, their technique matched the accuracy of dense-communication baselines while using lower bandwidth and achieving faster convergence.

Additionally, Aji and Heafield [47] introduced a simple yet effective technique for sparsifying gradient updates, supporting quantization across a broad domain. They discovered that gradient updates were heavily skewed toward zero, allowing only 1% of learning dynamics changes to be preserved with minimal information loss, albeit at a slower convergence rate. The work of Seide et al. [48] further advanced this field by evaluating 1-bit gradient quantization with error feedback in distributed SGD for speech-related data. Their method achieved up to 10× speedups recorded by training hundreds of hours of speech data in few hours with negligible accuracy loss. This demonstrated that even extreme quantization could maintain network convergence, with slight beneficial effects due to interaction with AdaGrad optimization technique.

3. Methodology

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to guarantee clarity and replicability. The approach utilized as shown in Figure 1 consists of four stages namely; search strategy, selection of studies, data extraction, and evaluation of quality.

3.1. Search strategy

The systematic review utilized an extensive search approach across prominent academic databases including IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and arXiv's computer

science archive (cs.LG). The search query integrated three major concept groups through Boolean operators: (1) federated learning terms ("federated learning" OR "distributed machine learning"), (2) efficiency goals ("communication efficiency" OR "bandwidth optimization"), and (3) technical methods ("quantization" OR "pruning" OR "sparsification"). The search focused on peer-reviewed research published from 2018 to 2024 to capture the latest developments in FL related studies. The inclusion criteria provided in Table 1 mandated that studies must explicitly focus on communication efficiency in FL systems by empirically evaluating techniques for compression or aggregation. Theoretical papers lacking implementation results, works that concentrated exclusively on centralized learning and publications not in English were removed. The criteria were also applied during title/abstract screening and full-text review.

Figure 1
The PRISMA model

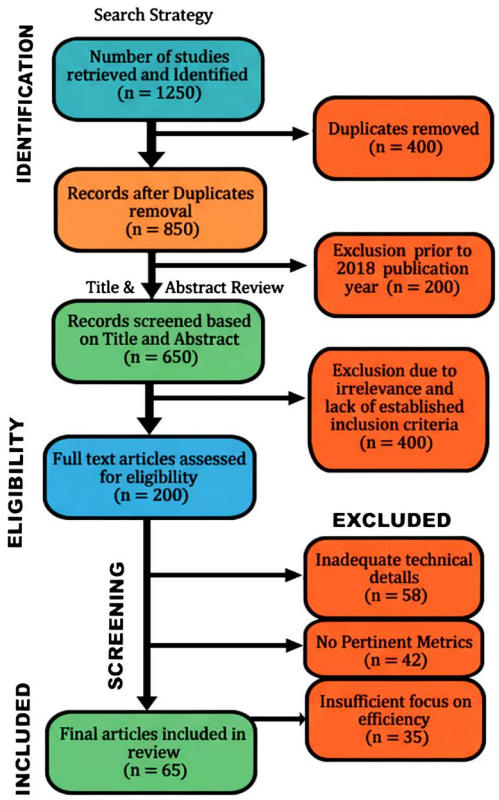


Table 1
Studies inclusion and exclusion criteria

S/nT	Category	Inclusion criteria	Exclusion criteria
1.	Publication date	Studies published between 2018–2024	Pre-2018 publications (prior to FL formalization by Konečný et al., 2016)
2.	Studies design	Empirical evaluations of FL systems	Theoretical papers without implementation results
3.	Technical focus	Explicit examination of communication efficiency (compression/ aggregation techniques)	Studies focusing solely on privacy/security without efficiency metrics
4.	Performance data	Reports quantitative metrics: bandwidth reduction, accuracy loss, or convergence rates	Missing baseline comparisons or insufficient performance data
5.	Accessibility	Full text available in English	Non-English papers without certified translations
6.	Publication type	Peer-reviewed conference papers/ journal articles	Grey literature (preprints without peer review, theses, white papers)

Table 2
Data extraction template with studies entries

S/n	Author	Techniques	Performance metrics	Strengths	Limitations	Implementation
1.	Bouacida et al. [31]	Adaptive Federated Dropout (AFD)	Convergence time, generalization accuracy	Reduces communication and computation costs, improves generalization	Potential trade-off between dropout rate and model performance	Empirical evaluation on standard tasks
2.	Chen et al. [49]	Asynchronous model update with temporally weighted aggregation	Communication cost, model accuracy	Enhances accuracy and convergence, reduces communication rounds	May require careful tuning of update frequencies	Empirical evaluation on two datasets
3.	Hu et al. [21]	Comprehensive survey	Discussion on common metrics (Communication rounds, model accuracy)	Comprehensive Overview	No empirical Evaluation	No Implementation
4.	Li et al. [22]	Local Differential Privacy, Client Selection	Communication efficiency, privacy preservation	Enhanced privacy with reduced communication	Potential computational overhead on clients	Tested on vehicle network datasets

Table 2 *Continued*

S/n	Author	Techniques	Performance metrics	Strengths	Limitations	Implementation
5.	Lu et al. [23]	Dynamic regularization	Model performance, Efficiency, convergence speed	Balances communication and learning performance	Computational complexity and Sensitive to data distribution	Federated learning simulation setup
6.	Abrahamyan et al. [10]	Learned Gradient Compression	Compression ratio, Accuracy	Highly efficient gradient transmission	Requires precise gradient correlation	Lightweight auto –encoder setup.
7.	Yujia et al. [50]	Adaptive communication strategies	Convergence rate, Accuracy	Adapts frequency of communication dynamically	Tuning hyperparameters may be complex	Python-based federated setup
8.	Yang et al. [37]	Model compression (FLCP framework)	Communication cost, Accuracy	Effective compression, preserves accuracy	Potential computational overhead on clients	Simulated experiments on standard datasets
9.	Zhao et al. [51]	Adaptive Quantized Gradient (AQG) with client dropout consideration	Accuracy, Communication rounds, Transmission reduction percentage, model convergence	Dynamically adjusts quantization level, Reduces transmission loss up to 50%, robust to up to 90% client dropout	Might degrade accuracy under high compression, Complexity in adaptive quantization level adjustment	Python simulation with federated setting
10.	Nguyen et al. [52]	Buffered Asynchronous Aggregation	Training speed, model accuracy	Handles client heterogeneity effectively	Complex buffer management	Tested on various FL scenarios
11.	Oh et al. [53]	Quantized Compressed Sensing	Accuracy, Compression ratio	Significant reduction in communication	Complex reconstruction, sensitive to noise	Tested on synthetic and real-world data
12.	Reisizadeh et al. [24]	Periodic Averaging, Quantization (FedPAQ)	Communication efficiency, convergence	Balances communication cost and model accuracy	Less effective in highly heterogeneous data and requires synchronization among clients	Simulations on federated datasets
13.	Wang et al. [36]	Adaptive Optimization (FedCAMS)	Convergence rate, communication rounds	Combines adaptivity with communication efficiency; theoretical convergence guarantees	Complexity in implementation; may require parameter tuning	Experiments on various benchmarks
14.	Yang et al. [54]	Over-the-Air Computation, Scheduling Policies	Bandwidth efficiency, latency	Improved communication in wireless networks	Susceptible to channel noise	Theoretical analysis and simulations
15.	Zhang et al. [55]	FedAvg with Compression, Asynchronous Updates	Convergence speed, scalability	Supports heterogeneous data and clients, improves scalability	Complexity in implementation, Requires synchronization mechanisms to handle asynchrony	Evauation on LEAF benchmark with up to 300 nodes
16.	Xu et al. [56]	Sparse Ternary Compression (STC)	Communication cost, model performance	Effective in bandwidth-constrained scenarios	May affect model convergence	Evaluated on standard FL tasks
17.	Sun et al. [57]	Gradient Compression	Communication overhead, model accuracy	Suitable for wireless-edge environments	Limited to specific network architectures	Simulations in wireless-edge settings
18.	Al-Saedi et al. [58]	Clustering optimization using Silhouette Index validation	Communication overhead, model accuracy	Reduces communication without sacrificing accuracy	Does not account for system heterogeneity	Implemented with dynamic worker clustering
19	Wu et al. [59]	Knowledge Distillation (FedKD)	Communication cost reduction; model accuracy	Significant reduction in communication cost, maintains competitive model performance	Potential computational overhead due to dual-model training	Implemented and evaluated on benchmark datasets

Table 3
Derived taxonomy for coding technical approaches

Field	Categories
Compression type	Quantization, Pruning, Sparsification, Knowledge distillation, Hybrid
Aggregation method	FedAvg, FedProx, SCAFFOLD, FedAdam, FedYogi, Custom
Dataset type	Image (CIFAR, FEMNIST), Text (Shakespeare), Synthetic (LEAF), Real-world

Figure 2
The PRISMA study selection process

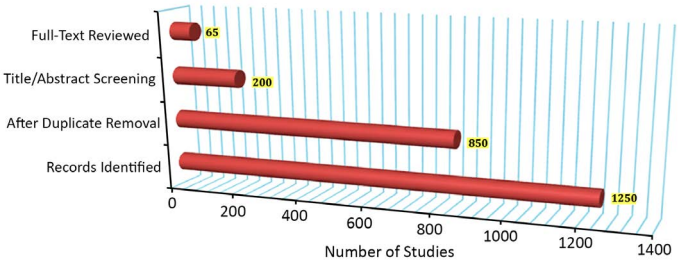


Table 4
Sampled studies analyzed based on derived taxonomy

S/n	Author	Compression type	Aggregation method	Dataset type
1.	Abbas et al. [12]	Pruning (model slimming)	FedAvg	Real-world (medical images)
2.	Jia et al. [33]	Hybrid (Sparsification + Quantization)	Custom	Synthetic
3.	Azzedin et al. [60]	Hybrid (pruning + quantization)	FedAvg, Custom	Real-world (banking loan dataset from Kaggle)
4.	Jia et al. [19]	Blockchain + Hybrid	Custom	Real-world
5.	Yang et al. [61]	Quantization	FedAvg	Real-World image data
6.	Berkani et al. [20]	Not available	Custom	Real-world
7.	Bouacida et al. [31]	Pruning (adaptive dropout)	FedAvg	Synthetic (LEAF datasets)
8.	Shen et al. [46]	Hybrid (Sparsification + Quantization)	Custom	Synthetic
9.	Jaggi et al. [42]	Coordinate pruning + Compression	FedAvg	Synthetic
10.	Lv et al. [62]	Quantization (over-the-air comp.)	FedAvg	Real-world (CIFAR 10)
11.	Sabahet al. [30]	Hybrid (quant + pruning + others)	FedAvg, FedAdam	Real-world
12.	Chen et al. [5]	Blockchain + Hybrid	Custom	Real-world
13	Chen et al. [40]	Pruning + Adaptive pruning	FedAvg	Real-world
14.	Chen et al. [63]	Sparsification	FedAvg	Synthetic
15.	Lu et al. [64]	Hybrid (adaptive async + pruning)	Custom	Real-world
16.	Chen et al. [49]	Knowledge distillation	FedAvg	Real-world (wearable healthcare)
17.	Crowley et al. [65]	Pruning	FedAvg	Synthetic
18.	Li et al. [38]	Hybrid	FedAvg	Real-world
19.	Pfeiffer et al. [4]	Quantization + Local SGD	FedAvg	Synthetic
20	Michalek et al. [18]	Federated learning protocol	FedAvg	Real-world (mobile keyboard)

Table 5
Sampled studies analyzed through the correlation of application and code reproducibility

S/n	Author	Compression type	Aggregation method	Application	Code reproducibility
1.	Sattler et al. [66]	Sparsification + Robust FL	FedAvg	Non-IID data	Not available
2.	Shaik et al. [67]	Hybrid (Stacked FL)	Personalized FL	Activity monitoring	Not available
3.	Sun et al. [57]	Gradient compression	FedAvg	Wireless uplink NOMA systems	Not available
4.	Suresh et al. [68]	Quantization	Mean estimation	Distributed learning	Not available
5.	Wang et al. [69]	Variable bitwidth quantization	FedAvg	Wireless networks	Not available
6.	Wang et al. [70]	Adaptive FL	Adaptive FL	Edge computing	Publicly available on arXiv preprint page
7.	XimiRng et al. [71]	Personalized FL	Adaptive personalized FL	Mobile edge environments	Not available
8.	Yao et al. [72]	Hybrid (Two-stream FL)	FedAvg	Communication cost reduction	Not available

S/n	Author	Compression type	Aggregation method	Application	Code reproducibility
9.	Zhou et al. [73]	Quantization (8-bit fixed point)	FedAvg	Sparse MobileNetV2	Not available
10.	Zhang et al. [74]	Adaptive aggregation (FedPD)	FedPD	Non-IID data	Not available
11.	Zhang et al. [25]	Representation learning	FedAvg	Human mobility prediction	Not available
12.	Aji and Heafield [47]	Hybrid (Adaptive computation & communication compression)	FedAvg	Federated edge learning	Not available
13.	Wang et al. [27]	None (Non-IID data)	FedDual-Decoupling	Non-IID data	Not available
14.	Stich et al. [32]	Sparsification + Memory optimization	FedAvg	Wireless edge devices	Data-driven study; code availability varies
15.	Zhou et al. [75]	Efficient + Fair FL	Personalized FL	Wearable devices	Not available
16.	Zhu et al. [11]	Analog aggregation	FedAvg	Federated edge learning	Not available
17.	Zhu et al. [76]	Pruning (Layer-wise pruning)	FedAvg	FL compression	Not available

3.2. Study selection process

The selection of studies adhered to the PRISMA 2020 guidelines [70] with model depicted in Figure 1, starting with 1250 records identified initially. Following the elimination of 400 duplicates via automated tools and manual checks, 850 studies were subjected to title and abstract review.

Two separate reviewers assessed each record based on the established inclusion criteria, resulting in high inter-rater reliability (κ = 0.82). This stage eliminated 650 studies that failed to satisfy methodological or topical criteria. The remaining 200 studies underwent full-text review, with a total of 135 studies being excluded because of inadequate technical detail (n = 58), absence of pertinent metrics (n = 42), or failure to obtain complete manuscripts (n = 35). The final dataset included 65 high-quality studies for extracting and analyzing data as shown in Figure 2.

3.3. Categorization criteria for academic and industrial studies

To systematically categorize studies as either academic or industrial, we applied the following criteria. A study was classified as industrial if it met any of these conditions: (1) at least one author was affiliated with a corporate or industry lab, (2) the study explicitly referenced industrial deployment, commercial applications, or internal production systems, or (3) the work was sponsored or conducted by an industrial research consortium.

Conversely, studies were labeled as academic only if all authors were affiliated with academic or non-profit research institutions, the research was exploratory or in a pre-deployment stage, and there was no mention of industry applications or funding. This classification framework enabled a clear and structured comparison of trends, methodologies, and priorities between academic research and industry practices.

3.4. Data extraction protocol

A standardized extraction template captured six key dimensions from each study: (1) bibliographic information (authors, publication year), (2) adopted technique (3) performance metrics (bandwidth reduction percentages, accuracy metrics, convergence rates), and (4) identified limitations (computational overhead, scalability constraints). The extraction process was conducted by two trained researchers using a pilot-tested Google Sheets form, with weekly reconciliation meetings to resolve discrepancies. For studies reporting multiple experiments, results from standardized benchmarks like LEAF or FEMNIST datasets to ensure comparability are prioritized in the selection. A sample of the extraction template is depicted in Table 2.

In order to have an in-depth analysis of the techniques adopted in the studies, a novel taxonomy shown in Table 3 was developed for coding technical analysis and further used in determining the suitability of the selected studies with knowledge derived for further analysis and reporting of the studies.

Findings from the sampled analysis according to the established taxonomy are presented in Table 4. Table 5 presents a depiction of the studies describing the type of comprehension, aggregation method, various application areas, and assesses the reproducibility status of the codes.

Studies included in Table 6 played an important role in the field of FL research. They covered various main areas such as benchmarking, setting standards, suggesting new system designs, and introducing new technical ideas like ways to make communication more efficient, better optimization techniques, privacy protection methods, and specific implementations for different fields.

The selection of these studies was based on their main purpose and the methods they used in accordance to PRISMA model [77]. Specifically, studies about benchmarking and general system design were included because they provided valuable insights into how FL systems work, tested them across different hardware setups, or offered ideas for future system development. These studies focused more on guiding the big picture of FL systems rather than fine-tuning individual parts.

In addition, studies in categories like compression, optimization, or privacy were chosen because they proposed or tested specific methods that help FL systems run more efficiently, reliably, or on a larger scale. Other studies that looked at FL in specific areas like healthcare or finance were included if they showed how FL can be used in real-life, impactful situations. These helped validate methods and showed how FL can be useful in different fields.

Some studies covered more than one area, like combining optimization with privacy or incentives. In such cases, each article was placed in the category that best reflected its main focus based on what it discussed and what its research goals were. This helps explain why certain studies are in one summary table and not another, especially when they lack the system testing information or design features that are typical in architecture or standardization-focused studies.

Ultimately, these tables served to transparently illustrate the diversity and depth of FL research included in the review, ensuring clarity in how each work contributes to the overall narrative, whether through foundational system proposals or implementation-level innovation.

Table 6
Sampled studies analyzed through benchmarking, standardization and high-level proposal

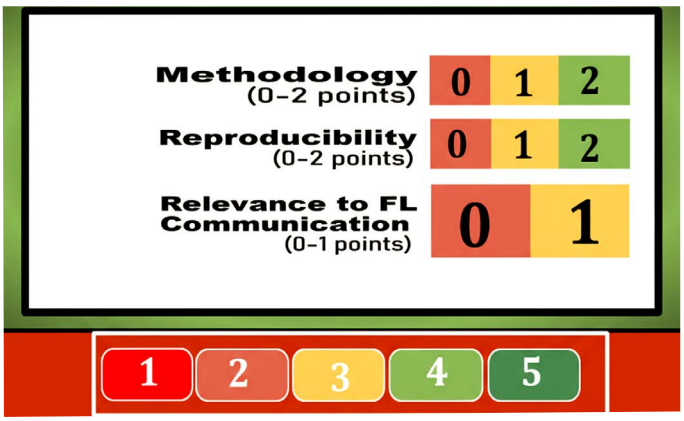
S/n	Author	Category	Focus area/Contribution
1.	Dehghani and Yazdanparast [1]	High-Level Architectural Proposal	Broad survey of distributed deep learning paradigms, highlighting system-level evolution from distributed ML to advanced FL architecturest
2.	Luoet al. [2]	Privacy & Data Handling	Privacy-preserving clustering for non-IID data in federated learning.
3.	Qayyum et al. [14]	Domain-Specific Application (Healthcare)	Multi-modal federated learning for COVID-19 diagnosis at the edge.
4.	Dhanawat et al. [16]	Domain-Specific Application (Finance)	Application of federated AI for financial risk management.
5.	Costanzo et al. [26]	Benchmarking	Evaluates SYCL performance portability across CPUs, GPUs, and hybrid systems.
6.	Ji et al. [28]	High-Level Architectural Proposal	Broad overview of trends in federated learning, including model fusion and Federated-X.
7.	Alistarh et al. [29]	Compression & Communication Efficiency	QSGD: gradient quantization for communication-efficient distributed SGD.
8.	Karimireddy et al. [34]	Optimization	SCAFFOLD: stochastic controlled averaging to address client drift in FL.
9.	Kang et al. [78]	Optimization & Incentive Mechanisms	Joint optimization using reputation and contract theory to incentivize reliable FL.

3.5. Quality assessment

Study quality was evaluated using an adapted 5-point QualSyst scale as shown in Figure 3 for assessing three domains. First, is the utilized methodology with clear description of FL architecture and baselines, reproducibility through the availability of code/ hyperparameter and relevance to communication efficiency (explicit focus on bandwidth or latency metrics). Each domain was scored independently by two reviewers with points ranging from 0 to 2 for methodology and reproducibility, while relevance to communication efficiency was scored on a scale of 0–1.

The aggregate score for each paper was recorded and 65 studies meeting the quality threshold scale between 4 and 5 points were selected. We documented excluded studies with their exclusion rationales in a supplemental file to maintain transparency. The quality assessment revealed that 78% of included studies provided open-source implementations, while only 42% reported detailed computational resource requirements which denote a notable gap in current reporting practices.

Figure 3
QualSyst scale rating scale



4. Results and Findings

The outcome from this systematic review depicts various results which are sub-divided into sections for clarity. The first section describes the study trends and characteristics followed by the quantitative analysis of performance benchmark of techniques. The interaction and gap between industry and academic trend in the adoption of FL was identified and lastly challenges to their implementation of FL were identified.

4.1. Geographical characteristic and study trend

The systematic analysis of the studies revealed distinct methodological and geographical patterns in communication-efficient FL research. The majority of studies amounting to 58.5% with a total number of 38 as shown in Figure 4 focused primarily on compression techniques, while aggregation optimization and hybrid with a count of 19 (29.2%) and 8 (12.3%) studies respectively represented smaller but growing segments. Geographically, North American institutions led contributions with a rating score of 42%, followed by Europe (31%) and Asia (27%) as shown in Figure 5.

Also, benchmark datasets displayed in Figure 6 showed clear dominance, with FEMNIST appearing in 43% of studies and CIFAR-10 in 34%, while real-world industry datasets accounted for only 6% of evaluations.

Figure 4
Analysis of compared FL techniques

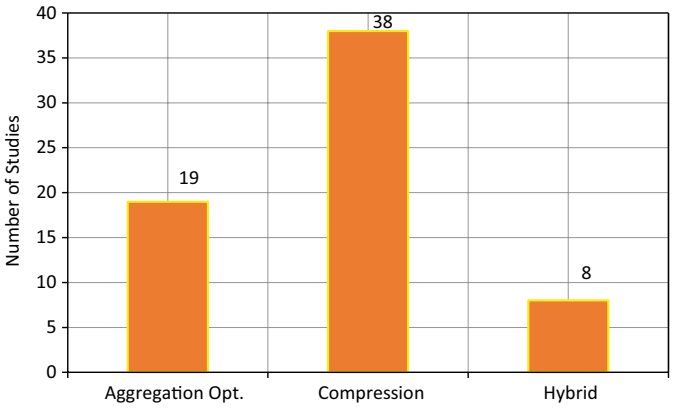


Figure 5
Geographical analysis of FL contribution

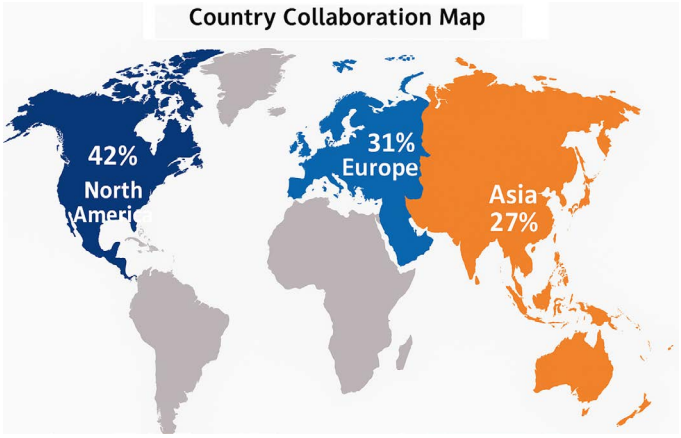
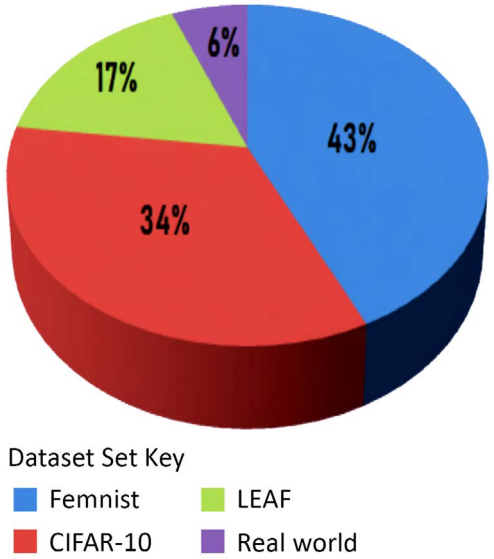


Figure 6
Dataset utilization across studies



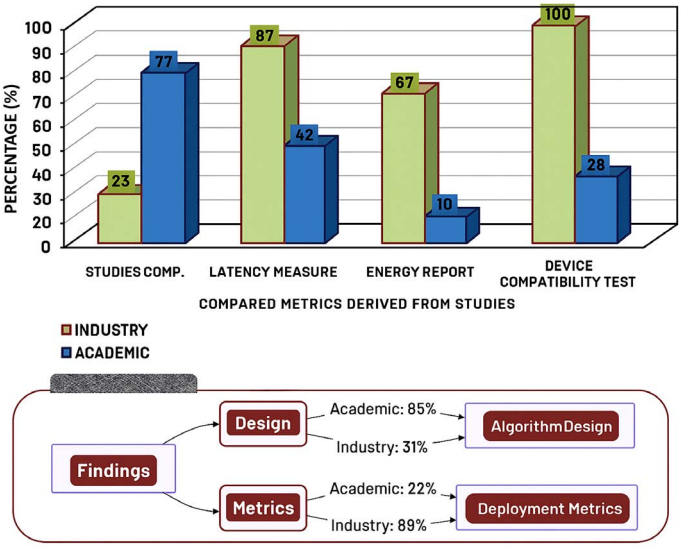
4.2. Performance analysis of FL techniques

Quantitative analysis demonstrated significant variation in technique effectiveness. As shown in Table 7, quantization methods achieved the highest bandwidth reduction (with a benchmarked mean value of 68.3% ± 12.1) but incurred greater convergence overhead (1420 ± 310 rounds) compared to pruning approaches. Notably, hybrid techniques combining 1-bit quantization with FedProx aggregation showed optimal balance, limiting accuracy degradation to a percentage level less than 2% points while maintaining 70%+ bandwidth savings.

Table 7
Comparative performance of FL optimization techniques

Technique category	Bandwidth reduction (%)	Accuracy Δ (pp)	Convergence rounds
1. Quantization (n = 32)	68.3 ± 12.1	−2.4 ± 1.8	1420 ± 310
2. Pruning (n = 21)	54.7 ± 9.6	−5.1 ± 2.3	980 ± 210
3. Hybrid (n = 12)	72.8 ± 8.4	−1.9 ± 1.2	1650 ± 290

Figure 7
Compared metrics derived from studies



4.3. Industry and academic research trend

A clear divergence emerged between academic and industry-led research trends in the analyzed corpus as shown in Figure 7. Industry research, which constituted 23% of the total studies, demonstrated a strong emphasis on practical deployment factors. Specifically, 87% of industry papers included latency measurements, a significant contrast to only 42% of academic papers that addressed this aspect. Additionally, energy efficiency was a key concern in the industry, with 67% of the studies reporting power consumption metrics, compared to just 10% of academic research. Industry efforts also prioritized cross-device compatibility, with all industry papers (100%) testing on three or more device types, whereas only 28% of academic studies did the same. In contrast, academic research predominantly focused on advancing theoretical innovations. A notable 92% of novel aggregation algorithms were developed within universities, underscoring academia's contribution to foundational advancements.

Furthermore, 78% of the studies addressing optimization for non-IID (non-independent and identically distributed) data originated from academic laboratories. These findings highlight the complementary roles of industry and academia, with the former concentrating on real-world application challenges and the latter driving theoretical progress.

Figure 8
FL implementation challenges score

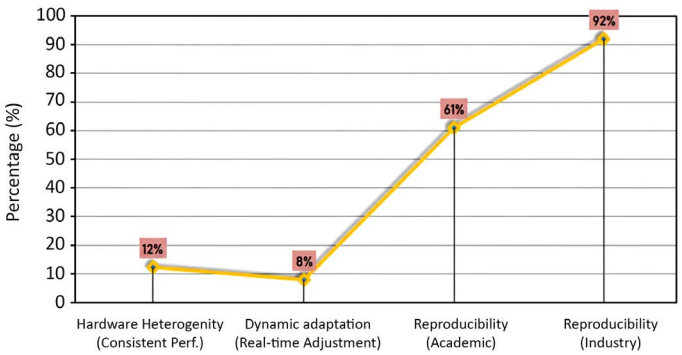
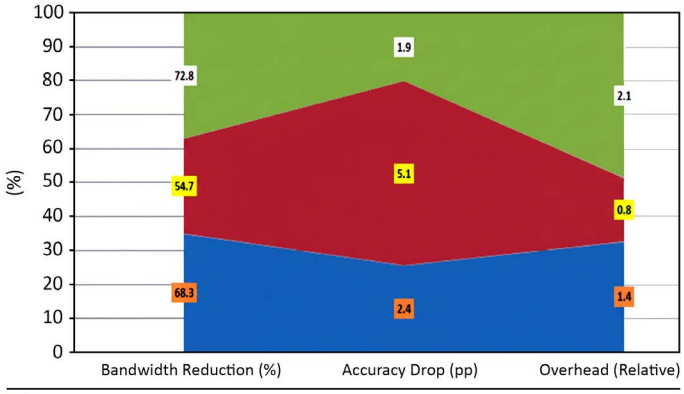


Figure 9
Trade-off chart among FL techniques



Hybrid	72.8	1.9	2.1
Pruning	54.7	5.1	0.8
Quantization	68.3	2.4	1.4

4.4. Implementation challenges

Implementation of methods faces several notable challenges that reoccur in different studies. First, hardware heterogeneity (differences) is a big problem. Only 12% of the methods evaluated could keep the same performance across different types of hardware like ARM, x86, and GPUs. This problem is a concern for real-world use, where models are expected to work smoothly on various edge and cloud systems.

Second, dynamic adaptation is rarely addressed as only 8% (five studies) made changes to compression settings on the fly to improve performance. This lack of flexibility makes it hard to scale FL solutions in environments where internet speed or computational capacity changes often.

Third, there is a big difference between what academic studies and industry research do when it comes to making things reproducible and ready for deployment. About 61% of academic studies share open-source code so others can try their work, but that goes up to 92% in industry research. However, even when studies are reproducible, there are often no standard tests, data sets, or clear ways to compare results, making it hard to validate findings.

Also, many academic studies only barely touch on or ignore issues like connecting with real-time data systems, working across different platforms and managing resources efficiently. These challenges show that reproducibility is not just about sharing code but also means being able to reproduce results in real-world settings. This includes having clear instructions, designing systems that are easy to use, and making sure they work across different hardware setups. Without these, moving from a simple idea to a large, working system is still haphazard and not well planned. Figure 8 gives an overall picture of these limitations, while Figure 9 shows the balance between saving bandwidth and keeping accuracy across different techniques.

5. Discussions

The findings of this systematic review reveal critical insights about communication efficiency in FL, with implications for both theoretical research and practical deployment. By analyzing 65 studies across academia and industry, we identify consistent patterns, unresolved challenges, and promising avenues for future work.

Figure 9 and Table 7 reveal that quantization-based methods are currently the best at reducing bandwidth, saving an average of 68.3%. This aligns with information theory principles, which suggest that using less precise data can still keep model accuracy acceptable. However, our analysis shows that there are more complex trade-offs when comparing quantization with other methods like pruning.

While 1-bit quantization reduces data transfer by up to 82% under ideal network conditions, it often requires 40–60% more training rounds compared to pruning-based methods. This longer training time can negate the initial bandwidth gains, especially in situations where speed and resources are limited. In contrast, pruning techniques (structured or layer-wise pruning) tend to converge more consistently and keep accuracy higher during early training, even though they offer lower average bandwidth savings (about 45–55%). These two methods also differ in computational overhead and how well they work with different models. Quantization usually fits well with existing training setups but may suffer from numerical instability in deep networks. Pruning, on the other hand, gives better control over gradients but often needs custom setup and retraining, which adds complexity.

Recent studies from 2023 to 2024, highlighted in the upper region of Figure 9, suggest hybrid approaches that try to balance these trade-offs. Combining adaptive quantization with robust aggregation methods like FedProx or SCAFFOLD has shown good results by saving 70% or more on bandwidth, keeping accuracy loss under 2% percentage points, and helping to reduce longer training times.

However, these methods can add extra computational and memory costs which may pose barriers to deployment on edge devices and mobile hardware. As quantization remains a dominant approach for improving communication efficiency, its benefits must be balanced against longer training times and higher system requirements. Pruning, though less effective at compression, offers more stable training and lower inference costs. The future lies in hybrid strategies that adjust based on deployment contexts including network conditions, device capabilities, and model complexity.

Considering the stark contrast between academic and industry priorities, it reveals a growing translational gap in FL research. Our data shows that 85% of university-led studies focus on novel algorithm design, particularly for non-IID data scenarios (78% of academic papers). In contrast, 89% of industry research prioritizes measurable deployment factors like latency spikes during model aggregation and energy consumption per training round — metrics reported in only 10% of academic papers (Section 3.3).

This divergence explains why just 17% of peer-reviewed techniques have been adopted in production systems, according to our analysis of open-source repositories. A representative example comes from Google's Gboard team who found that while academic 1-bit quantization methods reduced server costs by 40% [54], they increased client-side energy use by 35% — a trade-off rarely examined in theoretical papers. Figure 8 highlights several critical implementation challenges that frequently recur across studies and warrant immediate attention. First, hardware heterogeneity poses a significant barrier. While many current approaches are tested on uniform hardware, real-world edge networks often consist of a mix of ARM CPUs, GPUs, and TPUs.

In addition, only 12% of the reviewed methods demonstrated consistent efficiency across such heterogeneous hardware configurations (as discussed in Section 3.4). Second, dynamic network conditions present another major challenge. Most algorithms assume stable bandwidth, yet actual 5G/6G environments are characterized by rapid and unpredictable fluctuations. Notably, only five studies representing about 8% of the total incorporated real-time compression adjustment mechanisms, as shown in Figure 8.

Finally, energy efficiency remains a pressing concern. Aggressive compression strategies, as referenced in Table 2, can lead to a doubling of energy consumption, making them impractical for large-scale IoT deployments. This issue is further supported by findings in NVIDIA's recent whitepaper on FL in smart cities which highlights the unsustainable energy costs associated with such approaches.

Building on these insights, three key priority directions are proposed to advance the field in future. First, there is a pressing need

to replace static compression methods with adaptive compression techniques. Reinforcement learning has shown early potential in this area through improvement in the bandwidth-accuracy trade-off by dynamically adjusting bit-width during training. Second, hardware-aware FL co-design is essential to address energy constraints. The work by Kang et al. [78] on FPGA-accelerated aggregation highlights a promising path forward, particularly in mitigating the energy inefficiencies evident in the large-bubble clusters shown in Figure 3. Third, the field must establish standardized benchmarking practices. Current evaluations suffer from inconsistent metrics, with some studies measuring algorithm-level bandwidth savings and others assessing end-to-end system performance. A FL-specific extension of the machine learning performance (MLPerf) benchmark suite could provide a unified framework for fair comparison.

Additionally, regulatory bodies should require energy efficiency reporting in FL research similar to the EU's Ecodesign Directive to help bridge the gap between academic innovation and industry implementation. Although this study adhered to the rigorous PRISMA methodology, several limitations should be acknowledged.

A substantial portion of the reviewed literature (80%) was published after 2020, potentially overlooking foundational work in distributed optimization. More so, 39% of academic studies did not provide reproducible code, which is especially troubling given the applied nature of FL. These factors collectively suggest that caution is warranted when extending the study's conclusions to all FL contexts.

6. Conclusion

This systematic review has undertaken a comprehensive examination of communication-efficient FL through rigorous analysis of 65 peer-reviewed studies. The findings paint a nuanced picture of a field that has made significant theoretical advances, yet still faces substantial practical challenges in real-world implementation. At its core, our research confirms that there exists no universal solution for optimizing FL systems as each approach involves carefully considered trade-offs between bandwidth efficiency, computational overhead, and model accuracy.

The divergence between academic research and industry needs represents one of the most significant barriers to widespread FL adoption. While academic laboratories have made extraordinary progress in developing novel algorithms, these innovations often fail to address the practical constraints that dominate industry priorities. Our analysis shows that only very few of cutting-edge techniques transition from research papers to production systems, primarily due to insufficient attention to energy consumption, hardware compatibility, and dynamic network conditions. Looking forward, the field must address critical requirements to realize FL's full potential. First, we need adaptive compression frameworks that can automatically adjust their behavior based on real-time network conditions and device capabilities.

Second, the research community must prioritize hardware-aware algorithm design, moving beyond theoretical optimizations to consider actual processor architectures and energy constraints. Finally, and perhaps most urgently, the establishment of standardized evaluation benchmarks would enable meaningful comparison between approaches and accelerate progress toward deployable solutions. An FL-specific extension of the MLPerf framework could serve this purpose while maintaining backward compatibility with existing machine learning evaluation protocols.

For practitioners implementing FL systems today, our findings suggest several immediate recommendations such as hybrid approaches combining quantization with robust aggregation methods like FedProx currently offer the best balance of performance characteristics for most applications. Pilot testing should always be conducted on representative hardware configurations, as performance can vary dramatically across different processors and network conditions. Most importantly, teams must monitor both communication savings and local computation costs, optimizing one metric at the expense of the other often leads to suboptimal overall system performance.

7. Recommendation for Future Research Direction

As FL moves from being a research idea to a real-world tool used in businesses, the next big step in research needs to focus on making the best ideas work in practice. This will require unprecedented collaboration across traditionally separate domains including people who design algorithms working with engineers who build hardware, academic researchers need to collaborate with developers in industries, and decision-makers engaging with technical experts to create clear rules and frameworks.

To support this transition, the datasets and analysis frameworks developed through this systematic review have been made publicly available, serving as foundational tools to accelerate innovation and facilitate reproducible experimentation. However, as the field grows, there are still some important challenges and chances for more research.

First, there is an urgent need for standardized sustainability benchmarking in FL. Even though more people are interested in using it, not many studies look at how much energy it uses, how much carbon it produces, or how it affects resources across different networks. Future work should include ways to measure sustainability and create tools to check how well these systems save energy and improve communication.

Second, there is still a lack of open standards that make it easy for people to repeat results or use tools across different platforms. Creating shared sets of tests, places to store models (repositories), and ready-to-use toolkits will be key to bringing together different research efforts and making fair, clear comparisons between various methods.

Third, there is a lack of research on how well FL works in different areas outside of healthcare, finance, and the internet of things. Few studies narrowly scoped within healthcare, finance, or IoT, with limited exploration neglecting fields like education, farming, or government. Future work should explore how these methods can be adapted to fit the special needs and values of those areas, especially in places where computers and internet are not widely available.

Lastly, while communication efficiency has been studied a lot, there is not much work on making protocols that use energy wisely—especially for small devices that have limited power and might not always be connected. Emerging methods should focus on creating systems that use less time, are kinder to batteries, and can adjust to different levels of available resources, helping these systems work well on a wide range of devices and situations.

Conclusively, to get the most out of FL, we need more than just better algorithms. We also need smart system designs, teamwork across fields, and a focus on being green and inclusive. Tackling these many challenges will help create federated systems that are efficient, private, and can work everywhere while also being good for the planet and fair for all.

Acknowledgement

The authors acknowledge the efforts of the reviewers of this paper. We also appreciate their meaningful contribution, valuable suggestions, and comments to this paper which helped us in improving the quality of the manuscript.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Femi Temitope Johnson: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Elugbadebo Oladapo:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization, Supervision. **Olukumoro Olugbenga:** Software, Formal analysis, Resources, Data curation, Visualization, Supervision, Project administration. **Alomaja Victor:** Validation, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision. **Akande Adenike**: Methodology, Software, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Dehghani, M., & Yazdanparast, Z. (2023). From distributed machine to distributed deep learning: *A comprehensive survey. Journal of Big Data*, *10*(1), 158. https://doi.org/10.1186/s40537-023-00829-x

[2] Elhussein, A., & Gürsoy, G. (2023). Privacy-preserving patient clustering for personalized federated learnings. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, 150–166.

[3] Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H., & Dou, D. (2022). From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, *64*(4), 885–917. https://doi.org/10.1007/s10115-022-01664-x

[4] Pfeiffer, K., Rapp, M., Khalili, R., & Henkel, J. (2023). Federated learning for computationally constrained heterogeneous devices: A survey. *ACM Computing Surveys*, *55*(14s), 334. https://doi.org/10.1145/3596907

[5] Chen, L., Zhao, D., Tao, L., Wang, K., Qiao, S., Zeng, X., & Tan, C. W. (2025). A credible and fair federated learning framework based on blockchain. IEEE Transactions on Artificial Intelligence, 6(2), 301–316. https://doi.org/10.1109/TAI.2024.3355362

[6] Aggarwal, M., Khullar, V., & Goyal, N. (2024). A comprehensive review of federated learning: Methods, applications, and challenges in privacy-preserving collaborative model training. In J. Singh, S. Goyal, R. Kumar Kaushal, N. Kumar, & S. Singh Sehra (Eds.), *Applied data science and smart systems* (pp. 570–575). CRC Press. https://doi.org/10.1201/9781003471059-73

[7] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., & He, B. (2023). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3347–3366. https://doi.org/10.1109/TKDE.2021.3124599

[8] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 12. https://doi.org/10.1145/3298981

[9] Ayeelyan, J., Utomo, S., Rouniyar, A., Hsu, H. C., & Hsiung, P. A. (2024). Federated learning design and functional models: Survey. *Artificial Intelligence Review*, *58*(1), 21. https://doi.org/10.1007/s10462-024-10969-y

[10] Abrahamyan, L., Chen, Y., Bekoulis, G., & Deligiannis, N. (2022). Learned gradient compression for distributed deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 7330–7344. https://doi.org/10.1109/TNNLS.2021.3084806

[11] Zhu, G., Wang, Y., & Huang, K. (2020). Broadband analog aggregation for low-latency federated edge learning. *IEEE Transactions on Wireless Communications, 19*(1), 491–506. https://doi.org/10.1109/TWC.2019.2946245

[12] Abbas, Q., Daadaa, Y., Rashid, U., & Ibrahim, M. (2023). Assist-dermo: A lightweight separable vision transformer model for multiclass skin lesion classification. *Diagnostics*, *13*(15), 2531. https://doi.org/10.3390/diagnostics13152531

[13] Upreti, D., Yang, E., Kim, H., & Seo, C. (2024). A comprehensive survey on federated learning in the healthcare area: Concept and applications. *Computer Modeling in Engineering & Sciences, 140*(3), 2239–2274. https://doi.org/10.32604/cmes.2024.048932

[14] Qayyum, A., Ahmad, K., Ahsan, M. A., Al-Fuqaha, A., & Qadir, J. (2022). Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, *3*, 172–184. https://doi.org/10.1109/OJCS.2022.3206407

[15] Hernandez-Cruz, N., Saha, P., Sarker, M. M. K., & Noble, J. A. (2024). Review of federated learning and machine learning-based methods for medical image analysis. *Big Data and Cognitive Computing*, *8*(9), 99. https://doi.org/10.3390/bdcc8090099

[16] Dhanawat, V., Shinde, V., Karande, V., & Singhal, K. (2024). Enhancing financial risk management with federated AI. In *2024 8th SLAAI International Conference on Artificial Intelligence*, 1–6. https://doi.org/10.1109/SLAAI-ICAI63667.2024.10844982

[17] Xiao, B., Yu, X., Ni, W., Wang, X., & Poor, H. V. (2025). Over-the-air federated learning: Status quo, open challenges, and future directions. *Fundamental Research*, *5*(4), 1710–1724. https://doi.org/10.1016/j.fmre.2024.01.011

[18] Michalek, J., Oujezsky, V., Holik, M., & Skorpil, V. (2024). A proposal for a federated learning protocol for mobile and management systems. *Applied Sciences*, *14*(1), 101. https://doi.org/10.3390/app14010101

[19] Jia, Y., Xiong, L., Fan, Y., Liang, W., Xiong, N., & Xiao, F. (2024). Blockchain-based privacy-preserving multi-tasks federated learning framework. *Connection Science*, *36*(1), 2299103. https://doi.org/10.1080/09540091.2023.2299103

[20] Berkani, M. R. A., Chouchane, A., Himeur, Y., Ouamane, A., Miniaoui, S., Atalla, S., ..., & Al-Ahmad, H. (2025). Advances in federated learning: Applications and challenges in smart building environments and beyond. *Computers*, *14*(4), 124. https://doi.org/10.3390/computers14040124

[21] Hu, K., Li, Y., Xia, M., Wu, J., Lu, M., Zhang, S., & Weng, L. (2021). Federated learning: A distributed shared machine learning method. *Complexity*, *2021*(1), 8261663. https://doi.org/10.1155/2021/8261663

[22] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[23] Lu, C., Ma, W., Wang, R., Deng, S., & Wu, Y. (2023). Federated learning based on stratified sampling and regularization. *Complex & Intelligent Systems*, *9*(2), 2081–2099. https://doi.org/10.1007/s40747-022-00895-3

[24] Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., & Pedarsani, R. (2020). FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, *108*, 2021–2031.

[25] Zhang, X., Wang, Q., Ye, Z., Ying, H., & Yu, D. (2023). Federated representation learning with data heterogeneity for human mobility prediction. *IEEE Transactions on Intelligent Transportation Systems, 24*(6), 6111–6122. https://doi.org/10.1109/TITS.2023.3252029

[26] Costanzo, M., Rucci, E., García-Sánchez, C., Naiouf, M., & Prieto-Matías, M. (2025). Analyzing the performance portability of SYCL across CPUs, GPUs, and hybrid systems with SW sequence alignment. *Future Generation Computer Systems*, 170, 107838. https://doi.org/10.1016/j.future.2025.107838

[27] Wang, Z., Li, H., Li, J., Hu, R., & Wang, B. (2024). Federated learning on non-IID and long-tailed data via dual-decoupling. *Frontiers of Information Technology & Electronic Engineering*, *25*(5), 728–741. https://doi.org/10.1631/FITEE.2300284

[28] Ji, S., Tan, Y., Saravirta, T., Yang, Z., Liu, Y., Vasankari, L., ..., & Walid, A. (2024). Emerging trends in federated learning: From model fusion to federated X learning. *International Journal of Machine Learning and Cybernetics*, *15*(9), 3769–3790. https://doi.org/10.1007/s13042-024-02119-1

[29] Alistarh, D., Grubic, D., Li, J., Tomioka, R., & Vojnović, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1709–1720.

[30] Sabah, F., Chen, Y., Yang, Z., Raheem, A., Azam, M., Ahmad, N., & Sarwar, R. (2025). Communication optimization techniques in Personalized Federated Learning: *Applications, challenges and future directions. Information Fusion*, *117*, 102834. https://doi.org/10.1016/j.inffus.2024.102834

[31] Bouacida, N., Hou, J., Zang, H., & Liu, X. (2021). Adaptive federated dropout: Improving communication efficiency and generalization for federated learning. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops*, 1–6. https://doi.org/10.1109/INFOCOMWKSHPS51825.2021.9484526

[32] Stich, S. U., Cordonnier, J.-B., & Jaggi, M. (2018). Sparsified SGD with memory. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 4452–4463.

[33] Jia, J., Liu, J., Zhou, C., Tian, H., Dong, M., & Dou, D. (2024). Efficient asynchronous federated learning with sparsification and quantization. Concurrency and Computation: *Practice and Experience*, *36*(9), e8002. https://doi.org/10.1002/cpe.8002

[34] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, *119*, 5132–5143.

[35] Sattler, F., Wiedemann, S., Muller, K.-R., & Samek, W. (2019). Sparse binary compression: Towards distributed deep learning with minimal communication. In *2019 International Joint Conference on Neural Networks,* 1–8. https://doi.org/10.1109/IJCNN.2019.8852172

[36] Wang, H.-P., Stich, S. U., He, Y., & Fritz, M. (2022). ProgFed: Effective, communication, and computation efficient federated learning by progressive training. In *Proceedings of the 39th International Conference on MachineLearning*,1–21. https://doi.org/10.60882/CISPA.24614364.V2

[37] Yang, W., Yang, Y., Xi, Y., Zhang, H., & Xiang, W. (2024). FLCP: Federated learning framework with communication-efficient and privacy-preserving. *Applied Intelligence*, *54*(9–10), 6816–6835. https://doi.org/10.1007/s10489-024-05521-y

[38] Li, P., Cheng, G., Huang, X., Kang, J., Yu, R., Wu, Y., ..., & Niyato, D. (2023). Snowball: Energy efficient and accurate federated learning with coarse-to-fine compression over heterogeneous wireless edge devices. *IEEE Transactions on Wireless Communications*, *22*(10), 6778–6792. https://doi.org/10.1109/TWC.2023.3245601

[39] Woisetschläger, H., Erben, A., Wang, S., Mayer, R., & Jacobsen, H.-A. (2024). A survey on efficient federated learning methods for foundation model training. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8317–8325. https://doi.org/10.24963/ijcai.2024/919

[40] Chen, Z., Yi, W., Lambotharan, S., & Nallanathan, A. (2023). Efficient wireless federated learning with adaptive model pruning. In *2023 IEEE Global Communications Conference*, 7592–7597. https://doi.org/10.1109/GLOBECOM54140.2023.10437211

[41] Daumé, H., Phillips, J. M., Saha, A., & Venkatasubramanian, S. (2012). Protocols for learning classifiers on distributed data. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, *22*, 282–290.

[42] Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., & Jordan, M. I. (2014). Communication-efficient distributed dual coordinate ascent. In *28th Annual Conference on Neural Information Processing Systems*, 3068–3076.

[43] Yuan, K., Ling, Q., & Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, *26*(3), 1835–1854. https://doi.org/10.1137/130943170

[44] Agarwal, N., Suresh, A. T., Yu, F. X., Kumar, S., & McMahan, H. B. (2018). cpSGD: Communication-efficient and differentially-private distributed SGD. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7575–7586.

[45] Li, M., Andersen, D. G., Smola, A., & Yu, K. (2014). Communication efficient distributed machine learning with the parameter server. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, *1*, 19–27.

[46] Shen, Z., Mokhtari, A., Zhou, T., Zhao, P., & Qian, H. (2018). Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication. In *Proceedings of the 35th International Conference on Machine Learning*, *80*, 4624–4633.

[47] Aji, A. F., & Heafield, K. (2017). Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 440–445. https://doi.org/10.18653/v1/D17-1045

[48] Seide, F., Fu, H., Droppo, J., Li, G., & Yu, D. (2014). 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. *In Interspeech 2014*, 1058–1062. https://doi.org/10.21437/Interspeech.2014-274

[49] Chen, Y., Sun, X., & Jin, Y. (2020). Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation. *IEEE Transactions on Neural Networks and Learning Systems, 31*(10), 4229–4238. https://doi.org/10.1109/TNNLS.2019.2953131

[50] Wang, Y., Lin, L., & Chen, J. (2022). Communication-efficient adaptive federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, 22802–22838.

[51] Zhao, Z., Mao, Y., Shi, Z., Liu, Y., Lan, T., Ding, W., & Zhang, X.-P. (2024). AQUILA: Communication efficient federated learning with adaptive quantization in device selection strategy. *IEEE Transactions on Mobile Computing*, *23*(6), 7363–7376. https://doi.org/10.1109/TMC.2023.3332901

[52] Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Vincent Poor, H. (2021). Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 23(*3), 1622–1658. https://doi.org/10.1109/COMST.2021.3075439

[53] Oh, Y., Lee, N., Jeon, Y.-S., & Poor, H. V. (2023). Communication-efficient federated learning via quantized compressed sensing. *IEEE Transactions on Wireless Communications*, *22*(2), 1087–1100. https://doi.org/10.1109/TWC.2022.3201207

[54] Yang, K., Jiang, T., Shi, Y., & Ding, Z. (2020). Federated learning via over-the-air computation. *IEEE Transactions on Wireless Communications, 19*(3), 2022–2035. https://doi.org/10.1109/TWC.2019.2961673

[55] Zhang, Z., Gao, Z., Guo, Y., & Gong, Y. (2025). Heterogeneity-aware cooperative federated edge learning with adaptive computation and communication compression. *IEEE Transactions on Mobile Computing*, *24*(3), 2073–2084. https://doi.org/10.1109/TMC.2024.3492916

[56] Xu, J., Du, W., Jin, Y., He, W., & Cheng, R. (2022). Ternary compression for communication-efficient federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(3), 1162–1176. https://doi.org/10.1109/TNNLS.2020.3041185

[57] Sun, H., Ma, X., & Hu, R. Q. (2020). Adaptive federated learning with gradient compression in uplink NOMA. *IEEE Transactions on Vehicular Technology*, *69*(12), 16325–16329. https://doi.org/10.1109/TVT.2020.3027306

[58] Al-Saedi, A. A., Boeva, V., & Casalicchio, E. (2022). FedCO: Communication-efficient federated learning via clustering optimization. Future Internet, 14(12), 377. https://doi.org/10.3390/fi14120377

[59] Wu, C., Wu, F., Lyu, L., Huang, Y., & Xie, X. (2022). Communication-efficient federated learning via knowledge distillation. *Nature Communications*, *13*(1), 2032. https://doi.org/10.1038/s41467-022-29763-x

[60] Azzedin, F., Ghaleb, M., El-Alfy, Y., Katib, R., & Hossain, M. (2023). A federated learning approach to banking loan decisions. In *2023 International Symposium on Networks, Computers and Communications*, 1–7. https://doi.org/10.1109/ISNCC58260.2023.10323875

[61] Yang, N., Wang, S., Chen, M., Brinton, C. G., Yin, C., Saad, W., & Cui, S. (2022). Model-based reinforcement learning for quantized federated learning performance optimization. In *2022 IEEE Global Communications Conference*, 5063–5068. https://doi.org/10.1109/GLOBECOM48099.2022.10001466

[62] Lv, Y., Ding, H., Wu, H., Zhao, Y., & Zhang, L. (2023). FedRDS: Federated learning on non-IID data via regularization and data sharing. *Applied Sciences*, *13*(23), 12962. https://doi.org/10.3390/app132312962

[63] Chen, X., Pan, R., Wang, X., Tian, F., & Tsui, C.-Y. (2023). Late breaking results: Weight decay is all you need for neural network sparsification. In *2023 60th ACM/IEEE Design Automation Conference*, 1–2. https://doi.org/10.1109/DAC56929.2023.10247950

[64] Lu, R., Zhang, W., Li, Q., He, H., Zhong, X., Yang, H., ... & Alazab, M. (2024). Adaptive asynchronous federated learning. *Future Generation Computer Systems*, *152*, 193–206. https://doi.org/10.1016/j.future.2023.11.001

[65] Crowley, E. J., Gray, G., Turner, J., & Storkey, A. (2021). Substituting convolutions for neural network compression. *IEEE Access*, *9*, 83199–83213. https://doi.org/10.1109/ACCESS.2021.3086321

[66] Sattler, F., Wiedemann, S., Müller, K. R., & Samek, W. (2019). Robust and communication-efficient federated learning from non-iid data. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(9), 3400–3413. https://doi.org/10.1109/TNNLS.2019.2944481

[67] Shaik, T., Tao, X., Higgins, N., Gururajan, R., Li, Y., Zhou, X., & Acharya, U. R. (2022). FedStack: Personalized activity monitoring using stacked federated learning. *Knowledge-Based Systems*, *257*, 109929. https://doi.org/10.1016/j.knosys.2022.109929

[68] Suresh, A. T., Yu, F. X., Kumar, S., & McMahan, H. B. (2017). Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning, 70*, 3329–3337.

[69] Wang, S., Chen, M., Brinton, C. G., Yin, C., Saad, W., & Cui, S. (2024). Performance optimization for variable bitwidth federated learning in wireless networks. *IEEE Transactions on Wireless Communications*, *23*(3), 2340–2356. https://doi.org/10.1109/TWC.2023.3297790

[70] Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., & Chan, K. (2019). Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, *37*(6), 1205–1221. https://doi.org/10.1109/JSAC.2019.2904348

[71] Ximing, C., Xilong, H., Du, C., Tiejun, W., Qingyu, T., Rongrong, C., & Jing, Q. (2025). FedMEM: Adaptive personalized federated learning framework for heterogeneous mobile edge environments. *International Journal of Computational Intelligence Systems*, *18*(1), 84. https://doi.org/10.1007/s44196-025-00814-7

[72] Yao, X., Huang, C., & Sun, L. (2018). Two-stream federated learning: Reduce the communication costs. In *2018 IEEE Visual Communications and Image Processing*, 1–4. https://doi.org/10.1109/VCIP.2018.8698609

[73] Park, J. H., Kim, K. M., & Lee, S. (2022). Quantized sparse training: A unified trainable framework for joint pruning and quantization in DNNs. *ACM Transactions on Embedded Computing Systems (TECS)*, *21*(5), 1–22. https://doi.org/10.1145/3524066

[74] Zhang, X., Hong, M., Dhople, S., Yin, W., & Liu, Y. (2021). FedPD: A federated learning framework with adaptivity to non-IID data. *IEEE Transactions on Signal Processing*, *69*, 6055–6070. https://doi.org/10.1109/TSP.2021.3115952

[75] Zhou, P., Xu, H., Lee, L. H., Fang, P., & Hui, P. (2022). Are you left out? : An efficient and fair federated learning for personalized profiles on wearable devices of inferior networking conditions. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(2), 91. https://doi.org/10.1145/3534585

[76] Zhu, Z., Shi, Y., Luo, J., Wang, F., Peng, C., Fan, P., & Letaief, K. B. (2023). Fedlp: Layer-wise pruning mechanism for communication-computation efficient federated learning. In I*CC 2023-IEEE International Conference on Communications*, 1250–1255. https://doi.org/10.1109/ICC45041.2023.10278563

[77] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ..., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. https://doi.org/10.1136/bmj.n71

[78] Kang, J., Xiong, Z., Niyato, D., Xie, S., & Zhang, J. (2019). Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journa*l, *6*(6), 10700–10714. https://doi.org/10.1109/JIOT.2019.2940820

How to Cite: Johnson, F. T., Oladapo, E., Olugbenga, O., Victor, A., & Adenike, A. (2025). Communication-Efficient Federated Learning: A Systematic Review of Model Compression and Aggregation Techniques. *Artificial Intelligence and Applications*. http://doi.org/10.47852/bonviewAIA52026275