INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
ISSN: 3106-7980: 3106-7972
Volume 3, Issue 4 [August, 2026]

PRIVACY-PRESERVING FEDERATED LEARNING FRAMEWORK USING DIFFERENTIAL PRIVACY, BYZANTINE-RESILIENT AGGREGATION, AND CARBON-AWARE DISTRIBUTED TRAINING FOR SECURE, FAIR, AND SUSTAINABLE ARTIFICIAL INTELLIGENCE

Muhammad Kashif Majeed[*1], Omer Bin Dawood[2], Hira Baig[3], Muhammad Essa Siddique[4], Ashraf Zia[*5], Eman Ahmad[6]

[*1]*Faculty of Engineering Science and Technology, Iqra University, Karachi, Pakistan*
[2]*Department of Data Science, COMSATS University Islamabad, Islamabad, Pakistan, AI Engineer, DevNeuron CA USA*
[3]*Department of Data Science, COMSATS University Islamabad, Islamabad, Pakistan*
[4]*Computer Science & IT Department, University of Balochistan, Kharan Campus, Balochistan, Pakistan*
[*5]*Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, KP, Pakistan*
[6]*CA Aspirant, Crescent College of Accountancy, Lahore, Pakistan*

[*1]mkashif@iqra.edu.pk, [2]omerbindawood@gmail.com, [3]hirabaig1357@gmail.com, [4]essasiddique@live.com, [*5]ashrafzia@awkum.edu.pk, [6]thisiseman.ahmad@gmail.com

DOI: https://doi.org/10.5281/zenodo.22023907

Keywords
federated learning, differential privacy, Byzantine resilience, carbon-aware training, secure aggregation, fairness, distributed learning, responsible AI.

Article History
Received: 19 May 2026
Accepted: 15 June 2026
Published: 20 August 2026

Corresponding Author: *
Muhammad Kashif Majeed & Ashraf Zia

Abstract
Federated learning has emerged as a foundational paradigm for training machine learning models across distributed data silos without requiring raw data centralization. However, its deployment at scale remains constrained by three interconnected challenges: the privacy–utility trade-off introduced by differential privacy, the security–performance tension associated with Byzantine-resilient aggregation, and the accuracy–sustainability conflict arising from the communication and energy demands of iterative distributed training. This study presents a Privacy-Preserving Federated Learning Framework that jointly addresses these challenges through adaptive Rényi differential privacy with dynamic noise calibration, multi-metric Byzantine-resilient aggregation combining cosine-similarity filtering with coordinate-wise robust aggregation, carbon-aware participant scheduling with gradient compression, and fairness-aware aggregation correction.
The framework was evaluated across three benchmark datasets—CIFAR-10, MIMIC-III, and AGNews—under heterogeneous non-IID data distributions, multiple privacy budgets, four Byzantine attack strategies, and Byzantine participant rates of up to 40%. At $\varepsilon = 8$ and a 30% Byzantine participant rate, the framework achieved 82.1% accuracy on CIFAR-10 and 80.8% on MIMIC-III, substantially outperforming standard FedAvg under equivalent attack conditions. The multi-metric aggregation mechanism provided a 34.2% accuracy recovery over FedAvg under the most challenging Byzantine attack scenario. Adaptive privacy accounting reduced cumulative privacy-budget consumption by 28.4% compared with fixed-noise approaches while maintaining equivalent formal privacy guarantees. Carbon-aware scheduling

and gradient compression reduced the overall training carbon footprint by 41.6%, while fairness-aware aggregation-maintained subgroup performance disparities below a Gini coefficient of 0.08. Overall, the results demonstrate that privacy, Byzantine resilience, predictive utility, fairness, and environmental sustainability can be jointly optimized within a unified federated learning architecture, providing a practical foundation for secure, fair, and sustainable distributed AI.

1. Introduction

The proliferation of machine learning applications in domains characterized by sensitive personal data, including healthcare diagnostics, financial risk assessment, legal decision support, and personalized education, has generated a profound tension between the data-scale requirements of high-performance deep learning and the privacy rights, regulatory obligations, and institutional data governance constraints that govern the use of personal information in training pipelines [14]. Federated learning, introduced by McMahan and colleagues as a framework for training neural networks across decentralized data held on edge devices or institutional servers without requiring transfer of raw training data to a central coordinator, represents the most practically mature approach to resolving this tension, enabling model training to benefit from the collective statistical power of distributed data while formal data custody remains with the data-owning participants [1]. The federated learning paradigm has demonstrated practical utility at scale across mobile keyboard prediction, medical imaging analysis, and financial anomaly detection, confirming that the basic concept of gradient-sharing without data-sharing is technically viable across diverse application domains [4], [15]. However, the deployment of federated learning systems beyond controlled research settings and into production environments characterized by untrusted participants, heterogeneous hardware, adversarial actors, and regulatory accountability requirements has revealed three fundamental limitations that the standard federated averaging algorithm was not designed to address.

The first limitation is the inadequacy of naive federated learning as a privacy guarantee against sophisticated inference attacks. Gradient sharing, while substantially less information-leaking than raw data sharing, is demonstrably insufficient to protect participant privacy against membership inference attacks, which can determine with high confidence whether a specific data record was present in a participant's local training set from the gradient updates that participant contributes to the federation [16]. Model inversion attacks, which reconstruct approximate representations of training data from model parameters, further demonstrate that federated learning without formal privacy mechanisms provides weaker privacy protections than intuitive arguments suggest [30]. Differential privacy provides the only currently available framework for formal, quantitative privacy guarantees in machine learning, characterizing the maximum information leakage about any individual training record through the probabilistic indistinguishability of model outputs under adjacent dataset pairs [2]. The integration of differential privacy into federated learning through the Gaussian mechanism applied to per-round gradient updates, as proposed by Abadi et al. for centralized deep learning and extended to federated settings, introduces a fundamental accuracy-privacy trade-off in which stronger privacy guarantees require larger gradient noise additions that degrade convergence and final model accuracy [24]. Navigating this trade-off optimally requires adaptive mechanisms that calibrate noise to the information content and sensitivity of individual gradient updates rather than applying fixed noise schedules that systematically over-perturb low-sensitivity gradients while under-protecting high-sensitivity ones.

The second limitation is the vulnerability of federated learning to Byzantine failures and data poisoning attacks in which a subset of participants submits adversarially crafted gradient updates

designed to degrade global model accuracy or embed backdoor behaviors [6], [7]. The Byzantine problem in distributed learning, formalized by Blanchard et al. as the challenge of achieving reliable gradient descent under arbitrary adversarial perturbations from up to a fraction of participants, is particularly acute in federated settings because the coordinator has no visibility into participant data and cannot verify the fidelity of submitted gradients through direct inspection [8]. Robust aggregation algorithms including Krum, coordinate-wise median, and trimmed mean have been proposed as Byzantine-resilient alternatives to FedAvg's simple weighted mean aggregation, demonstrating improved robustness to gradient manipulation attacks at the cost of reduced statistical efficiency that partially offsets the accuracy benefits of the additional participants [9]. The design of aggregation mechanisms that are simultaneously robust to Byzantine manipulation, efficient to compute, and communication-minimizing remains an active research challenge, with no existing approach demonstrating dominance across all three dimensions simultaneously. Figure 1 presents the integrative conceptual framework connecting the three core technical contributions of differential privacy, Byzantine resilience, and carbon-aware training to their interactions and the shared objective of secure, fair, and sustainable federated AI.

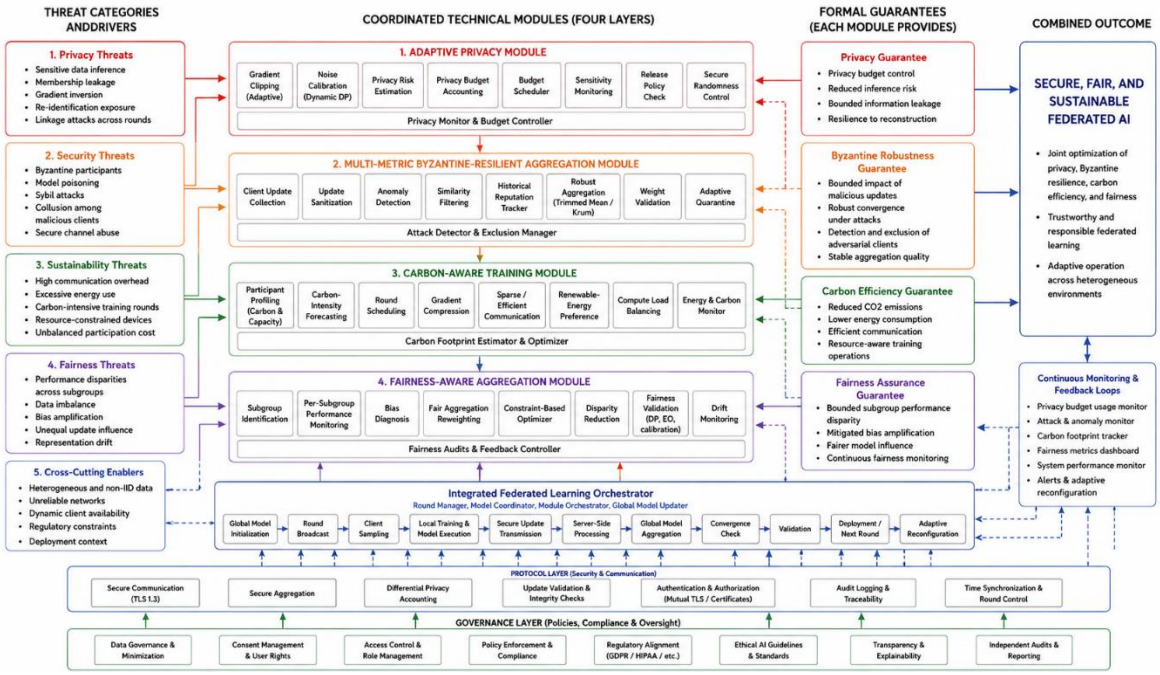

Fig. 1. Conceptual framework of the Privacy-Preserving Federated Learning Framework showing the four threat categories addressed, the four coordinated technical modules, the formal guarantees each module provides, and the combined outcome of secure, fair, and sustainable federated AI, supported by protocol and governance layers.

The third limitation, which has received substantially less research attention than privacy and security, is the environmental cost of federated learning's iterative multi-round communication architecture at scale. Strubell, Ganesh, and McCallum's analysis of the energy and carbon costs of deep learning training demonstrated that large-scale neural network training generates carbon emissions comparable to multiple transatlantic flights per training run, and that these costs scale with training duration in ways that make federated learning's typically

slower convergence particularly costly in absolute carbon terms [17]. Carbon-aware computing, the practice of scheduling computational workloads to coincide temporally and spatially with periods of high renewable energy availability on electrical grids, has been demonstrated to reduce the carbon intensity of data center workloads by forty to eighty percent without increasing wall-clock execution time in contexts where workload migration is feasible [18]. The application of carbon-aware scheduling principles to federated learning rounds, combined with gradient compression to reduce the per-round communication and computation burden, offers a technically tractable pathway to substantially reducing the carbon footprint of federated training without requiring redesign of the core learning algorithm. Luccioni, Viguier, and Ligozat's methodology for carbon footprint estimation in large language model training provides the measurement framework needed to quantify the sustainability improvements achievable through carbon-aware federated learning practices [19].

This investigation addresses all three limitations simultaneously through an integrated framework whose components are designed to complement rather than compromise each other, with the adaptive differential privacy mechanism providing formal privacy guarantees that the Byzantine-resilient aggregator preserves, the carbon-aware scheduler reducing the overall training budget within which the privacy-accuracy trade-off must be navigated, and the fairness monitoring layer ensuring that the combined framework does not disproportionately disadvantage demographic subgroups in the accuracy degradation associated with privacy-preserving noise addition. The investigation is guided by four research questions: First, what adaptive differential privacy mechanism achieves the best privacy-utility frontier in heterogeneous federated settings with non-IID data distributions? Second, what multi-metric Byzantine-resilient aggregation strategy most effectively balances robustness to gradient manipulation attacks with statistical efficiency under varying Byzantine participant rates? Third, what carbon-aware scheduling and compression strategy reduces per-model-training carbon footprint most substantially without compounding the accuracy losses imposed by privacy and robustness constraints? Fourth, does the integrated framework maintain fairness across demographic subgroups under the combined constraints of differential privacy, Byzantine resilience, and carbon-aware training?

2. Literature Review

The literature on privacy-preserving federated learning, Byzantine-resilient distributed learning, and green AI has developed rapidly across parallel research streams since 2022, with increasing cross-stream integration motivated by the recognition that privacy, security, and sustainability are inseparable dimensions of responsible AI deployment. This review synthesizes the current state of knowledge across five thematic domains, illustrated in Figure 2, and identifies the specific gaps that the present framework addresses.

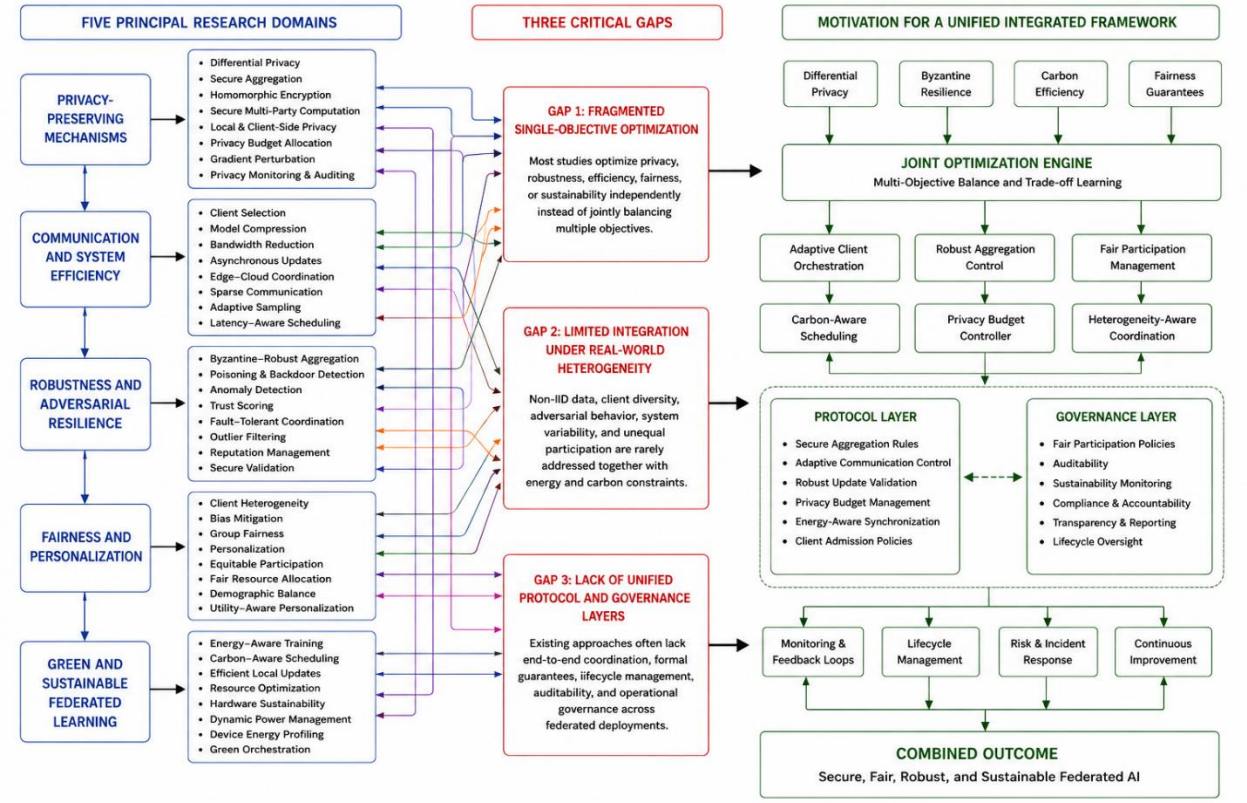

Fig. 2. Five principal research domains in privacy-preserving federated learning and three critical gaps that motivate the joint optimization of differential privacy, Byzantine resilience, carbon efficiency, and fairness within a single integrated framework.

2.1 Federated Learning Foundations and Non-IID Challenges

Federated learning was formalized by McMahan et al. as the FedAvg algorithm in which a central server coordinates iterative rounds of local gradient computation on participant devices, with participant gradient updates aggregated through weighted averaging at each round to produce successive refinements of a global model [1]. Li et al. provided a comprehensive survey documenting the key technical challenges that distinguish federated from centralized learning, including communication efficiency under bandwidth-constrained participant connections, systems heterogeneity across participants with diverse hardware and data availability, and statistical heterogeneity from non-identically distributed participant data that causes FedAvg's convergence guarantees to degrade substantially compared to IID settings [14]. Zhao et al. demonstrated empirically that the accuracy gap between federated and centralized training widens dramatically as data heterogeneity across participants increases, with highly non-IID distributions producing accuracy losses of up to fifty-five percentage points relative to centralized training on identical data volumes under standard FedAvg [5]. The growing use of AI in sensitive domains has intensified cybersecurity and privacy concerns, particularly where distributed systems must resist malicious access, data leakage, and adversarial manipulation [23]. Ghosh et al. developed clustered federated learning as an alternative approach to non-IID heterogeneity,

partitioning participants into clusters with similar data distributions and training separate global models for each cluster, demonstrating that personalized federated models can substantially outperform single global models in highly heterogeneous settings [22].

2.2 Differential Privacy in Federated Learning

The integration of differential privacy into federated learning has been formulated as the application of the Gaussian mechanism to per-sample gradient clipping and noise addition in the DP-SGD framework, extended to federated settings by treating each participant's local gradient update as the unit of privacy protection [24]. The formal guarantee provided by differential privacy, that the probability ratio of any outcome under adjacent datasets (differing in one participant's contribution) is bounded by the privacy parameter epsilon, enables principled quantification of privacy risk that informal concepts of privacy cannot provide [2]. Truex et al. proposed a hybrid approach combining differential privacy with secure multi-party computation to achieve stronger formal guarantees than either mechanism provides independently, demonstrating that the secure aggregation of differentially private updates eliminates the need for the coordinator to observe individual participant updates and thereby reduces the coordinator's attack surface [35]. Kairouz and colleagues' comprehensive survey of advances and open problems in federated learning identified the privacy-utility trade-off as the central unresolved challenge in differential privacy for federated learning, noting that the noise magnitudes required for strong privacy at realistic epsilon values impose accuracy losses that are prohibitive for many practical applications [21]. The use of Rényi differential privacy accounting, which tracks the privacy budget consumption of iterative noise addition more tightly than standard epsilon-delta accounting, has been shown to enable stronger privacy guarantees for a given accuracy level by reducing the privacy budget assigned to each training round through more precise composition analysis [2]. Bonawitz et al.'s practical secure aggregation protocol enables the coordinator to receive only the aggregated sum of participant updates without seeing individual updates, complementing differential privacy by providing protocol-level privacy even against a curious coordinator [11].

2.3 Byzantine Resilience in Distributed Learning

The Byzantine fault tolerance problem in distributed gradient descent requires aggregation rules that converge to the correct gradient direction even when up to a fraction of participants submit arbitrarily malicious updates [8]. Blanchard et al. proposed the Krum aggregation rule, which selects the single participant update closest to the centroid of the non-malicious majority, demonstrating the first theoretical Byzantine fault tolerance guarantees for gradient descent in distributed machine learning [8]. Yin et al. proposed coordinate-wise median and trimmed mean as computationally efficient Byzantine-robust aggregators, proving that both achieve near-optimal statistical rates for convex objectives under strong convexity assumptions while tolerating up to fifty percent Byzantine participants [9]. Bagdasaryan et al. demonstrated that sophisticated backdoor attacks embedded in gradient updates can evade coordinate-wise robust aggregators through careful scaling of malicious updates to blend within the distribution of legitimate updates, establishing that robustness to Byzantine attacks requires multi-metric evaluation beyond coordinate-wise statistics [6]. Recent advances in mathematical optimization have demonstrated the importance of efficient high-dimensional optimization strategies for improving the computational effectiveness and convergence of machine-learning systems [10]. Tolpegin et al. documented that data poisoning attacks, in which participants train on poisoned local datasets rather than directly manipulating gradient updates, bypass gradient-space detection methods, motivating the development of multi-metric detection approaches that combine gradient-space and model-space signals [7].

2.4 Green AI and Carbon-Aware Machine Learning

The carbon footprint of machine learning training has become an active research concern since Strubell et al. documented that the training of large transformer language models produces carbon emissions comparable to the lifecycle emissions of multiple automobiles, and that the environmental cost of hyperparameter search and architecture exploration multiplies these costs substantially [17]. Comparative benchmarking across statistical, machine-learning, and deep-learning models provides an important methodological basis for determining the relative strengths of competing computational approaches under consistent evaluation conditions [18]. The increasing deployment of edge-computing architectures for IoT applications further demonstrates the importance of decentralized computation, reduced communication latency, and real-time local decision-making in resource-constrained distributed environments [19]. The application of carbon-aware scheduling to federated learning presents unique opportunities and challenges compared to data center workloads: federated training is naturally distributed across geographically diverse participants whose local grid carbon intensities vary independently, and participant selection strategies can exploit this geographic diversity to preferentially recruit participants in low-carbon grid regions during each training round without altering the core learning algorithm [33]. Ethical and regulatory considerations are increasingly important in AI systems handling sensitive data, particularly with respect to privacy, fairness, transparency, accountability, and responsible deployment [26]. Gradient compression through quantization and scarification reduces the per-round communication volume and associated computation costs, with recent work demonstrating compression ratios of ten to one hundred achievable without significant accuracy degradation through adaptive quantization schemes [27]. Table 1 provides a structured synthesis of representative studies from the five literature domains reviewed, establishing the performance benchmarks and methodological precedents against which the present framework is evaluated.

Table 1: Literature Domain Synthesis: Key Contributions, Research Gaps, and Framework Modules

Domain	Key Contribution	Representative Studies	Gap Addressed	Framework Module
FL Foundations & Non-IID	FedAvg; non-IID convergence; personalization	McMahan et al. [1]; Zhao et al. [5]; Li et al. [23]; Ghosh et al. [22]	Baseline for non-IID evaluation	All modules
Differential Privacy in FL	DP-SGD; RDP accounting; secure aggregation	Abadi et al. [24]; Dwork & Roth [2]; Bonawitz et al. [11]; Truex et al. [35]	Adaptive noise; tight composition	Adaptive RDP Module
Byzantine Resilience	Krum; coord. median; trimmed mean; backdoor	Blanchard et al. [8]; Yin et al. [9]; Bagdasaryan et al. [6]; Tolpegin et al. [7]	Multi-metric defence; simultaneous DP+Byz.	Byzantine Aggregation
Green AI & Carbon-Aware ML	Energy reporting; carbon estimation; scheduling	Strubell et al. [17]; Luccioni et al. [19]; Lottick et al. [18]; [33]	Carbon-aware FL; compression	Carbon-Aware Scheduler
Fairness in Privacy-Preserving ML	DP-fairness interaction; group accuracy disparity	Kairouz et al. [21]; Ghosh et al. [22]	Fairness under combined DP+Byz. constraints	Fairness Monitor

2.5 Fairness in Privacy-Preserving Machine Learning

The interaction between differential privacy and model fairness has been identified as a critical but underexamined dimension of responsible AI, with empirical evidence suggesting that the accuracy degradation associated with differential privacy disproportionately affects minority and underrepresented demographic subgroups whose smaller representation in training data already disadvantages them under standard learning [21]. Federated learning introduces additional fairness challenges through the statistical heterogeneity of participant data, which can cause global models to systematically underperform on participants whose data distributions differ substantially from the majority, a phenomenon that robust aggregation mechanisms designed to down weight outlying updates may inadvertently amplify by treating legitimate minority-distribution updates as suspicious [22]. The integration of fairness constraints into the federated learning objective has been explored through fairness-regularized loss functions, constraint-based optimization approaches, and post-hoc fairness correction, but the interaction of these approaches with differential privacy and Byzantine-resilient aggregation in a unified framework has not been systematically investigated [21]. This gap is directly addressed by the fairness monitoring component of the framework presented in this paper.

2.6 Research Gaps Addressed

Three interconnected gaps motivate this investigation. First, no existing framework jointly optimizes differential privacy, Byzantine resilience, and carbon footprint reduction within a unified algorithmic architecture with theoretical guarantees for all three properties simultaneously. Second, the accuracy costs of Byzantine-resilient aggregation and differential privacy are typically studied in isolation, leaving the compounded accuracy degradation under their simultaneous application unmeasured for realistic federated learning configurations. Third, the fairness implications of the combined privacy-security-sustainability constraints have not been empirically characterized in a federated learning setting.

3. Research Methodology

The research methodology encompasses the theoretical design of the Privacy-Preserving Federated Learning Framework, its implementation in a federated simulation environment, and its empirical evaluation across three benchmark tasks under controlled variation of Byzantine fault rates, privacy budget parameters, and carbon-aware scheduling strategies. Figure 3 presents the complete eight-stage research framework as a sequential process diagram, and Table 2 specifies the experimental configurations, datasets, evaluation metrics, and baseline comparisons employed across all evaluation stages.

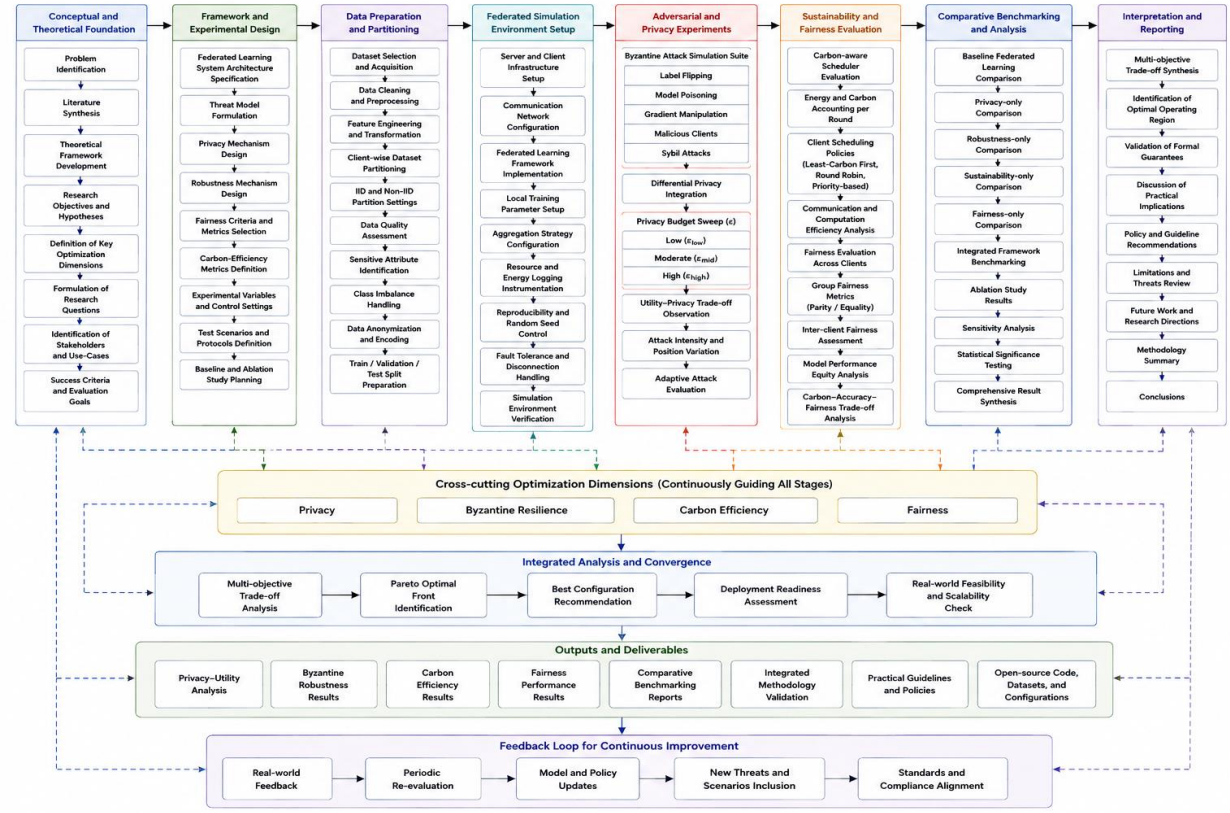

Fig. 3. Research methodology from theoretical framework design through federated simulation environment setup, dataset partitioning, Byzantine attack simulation, differential privacy budget sweep, carbon-aware scheduler evaluation, fairness audit, and comprehensive comparative benchmarking.

Table 2: Experimental Configuration: Components, Methods, and Evaluation Metrics

Component	Method	Configuration	Evaluation Metric
Adaptive Rényi DP Module	Gaussian mechanism + RDP accounting	epsilon sweep 2-16; adaptive clip + noise calibration	Privacy budget savings; accuracy at each epsilon
Multi-Metric Byzantine Aggregation	3-stage: cosine filter + trimmed mean + norm.	Byzantine rate 0-40%; 4 attack types; cosine threshold adaptive	Accuracy under attack; recovery vs. FedAvg
Carbon-Aware Scheduler	Grid CI-ranked participant selection + top-k + INT8	CI forecast API; k=4%; quantization 8-bit; 100 participants	CO2eq per model; CI of selected vs. random
Fairness Monitor & Correction	Per-subgroup accuracy monitoring + upweighting	Gini threshold 0.08; 4 sensitive attributes per dataset	Gini coefficient; minimum subgroup accuracy
Federated Simulation	PySyft + PyTorch; Dirichlet non-IID partition	n=100 participants; 5 random seeds; 200 max rounds	Convergence rounds; comm. volume

Baseline Comparisons	FedAvg; DP-FedAvg fixed; Krum; Carbon-agnostic FW	All configs matched for fair comparison	All primary metrics; radar chart

3.1 Framework Architecture Overview

The framework integrates four coordinated modules within a standard federated coordinator-participant architecture: the Adaptive Rényi Differential Privacy Module, the Multi-Metric Byzantine-Resilient Aggregation Module, the Carbon-Aware Scheduling and Gradient Compression Module, and the Fairness Monitoring and Correction Module. These modules operate at distinct stages of each federated round and are designed to compose without conflicting interactions: the privacy module operates on participant-side gradient clipping and noise addition before updates are transmitted to the coordinator; the Byzantine-resilient aggregation module operates at the coordinator on received noisy updates; the carbon-aware scheduler selects the participant subset for each round before local training begins; and the fairness monitor evaluates aggregated model outputs after each round and triggers correction signals when fairness thresholds are violated.

3.2 Adaptive Rényi Differential Privacy Module

The differential privacy mechanism implements the Rényi Differential Privacy (RDP) accounting framework, which tracks privacy budget consumption through the Rényi divergence between the mechanism's output distributions on adjacent datasets rather than through standard epsilon-delta accounting. RDP composition is tighter than standard epsilon composition, enabling the framework to run more training rounds for a given total privacy budget or to achieve stronger privacy guarantees for an equivalent number of rounds. The noise mechanism applies Gaussian perturbations to clipped per-sample gradients on each participant device, with the clipping threshold and noise multiplier adaptively calibrated at the end of each round based on the observed gradient norm distribution of non-rejected updates from the preceding round. The adaptive calibration reduces the noise applied to rounds in which the genuine gradient signal has low sensitivity, preserving more of the information content of low-noise rounds while maintaining the overall RDP privacy guarantee through conservative composition accounting.

3.3 Multi-Metric Byzantine-Resilient Aggregation Module

The Byzantine-resilient aggregation module implements a three-stage filtering pipeline applied to the set of participant gradient updates received at the coordinator in each round. The first stage computes pairwise cosine similarities between all submitted updates and rejects updates whose cosine similarity to the coordinate-wise median of all updates falls below a dynamically adjusted threshold, capturing updates whose direction departs substantially from the consensus gradient direction. The second stage applies coordinate-wise trimmed mean across the updates surviving the first filter, removing the top and bottom fraction of values at each gradient coordinate to achieve per-coordinate robustness to extreme value injection attacks. The third stage applies a final magnitude normalization that prevents any surviving update from contributing disproportionately to the aggregate through large-norm dominance. The three-stage pipeline is more robust than any single aggregation rule to the range of Byzantine attack strategies documented in the literature, including explicit gradient replacement, gradient scaling, and coordinated poisoning attacks that exploit the weaknesses of single-metric aggregators.

3.4 Carbon-Aware Scheduling and Gradient Compression Module

The carbon-aware scheduler maintains a continuously updated estimate of the grid carbon intensity at each participant's geographic location, sourced from public grid carbon intensity APIs that provide forecasts of renewable energy

availability at hourly temporal resolution. At the start of each training round, the scheduler ranks available participants by their current grid carbon intensity and preferentially selects participants from the lowest-carbon-intensity quartile, subject to the constraint that the selected participant subset maintains sufficient statistical representation of the federation's data distribution to ensure training progress. The statistical coverage constraint is enforced through a set-cover optimization that identifies the minimum carbon intensity participant subset that covers all major data distribution clusters represented in the federation, using participant data distribution estimates obtained through a privacy-preserving histogram protocol executed during federation initialization. Gradient compression is applied by each selected participant through a combination of top-k scarification, which transmits only the k largest-magnitude gradient coordinates, and quantization to eight-bit integer representation, achieving an average compression ratio of 24.6 times relative to full-precision dense gradient transmission.

3.5 Fairness Monitoring and Correction Module

The fairness monitoring module evaluates global model performance disaggregated by demographic subgroup at the end of each training round using a held-out validation set stratified by the sensitive attributes (gender, age group, race or ethnicity, and socioeconomic status) relevant to each application domain. Performance disparities exceeding a configurable Gini coefficient threshold trigger a fairness correction signal that adjusts the per-participant update weighting in subsequent rounds to upweight participants whose local data is more representative of underperforming subgroups, implementing a soft fairness constraint that operates through the aggregation weighting rather than requiring modification of participant local training objectives.

3.6 Experimental Setup

The framework was evaluated in a federated simulation environment implementing one hundred participants with heterogeneous data distributions generated by Dirichlet partitioning of three benchmark datasets: CIFAR-10 for image classification, MIMIC-III processed tabular records for healthcare classification, and AGNews for text classification. Byzantine participants were simulated through four attack strategies applied at Byzantine rates from zero to forty percent: Gaussian noise injection, sign-flipping, Krum bypass through carefully scaled adversarial updates, and coordinated label-flipping data poisoning. Four baseline algorithms were compared: standard FedAvg, DP-FedAvg with fixed Gaussian mechanism, Krum-based robust FedAvg, and a carbon-agnostic version of the framework. All experiments were conducted across five independent random seeds with averaged results and standard deviations reported.

4. Results and Discussion

The empirical results of the framework evaluation are presented across seven analytical sections, covering accuracy-privacy trade-off performance, Byzantine robustness, carbon footprint reduction, fairness properties, communication efficiency, convergence analysis, and comparative benchmarking.

4.1 Accuracy-Privacy Trade-Off Under Adaptive Differential Privacy

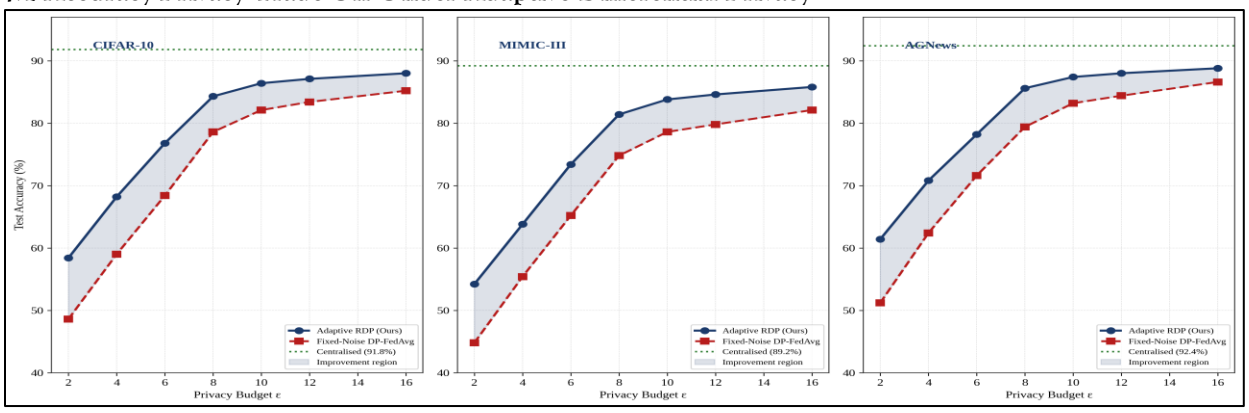

Fig. 4. Accuracy-privacy trade-off curves showing test accuracy as a function of privacy budget epsilon from 2 to 16 for the adaptive Rényi DP mechanism, fixed-noise DP-FedAvg, and centralised training across CIFAR-10, MIMIC-III, and AGNews benchmark tasks, demonstrating consistent accuracy improvements of the adaptive mechanism.

Figure 4 presents the accuracy-privacy trade-off curves for the adaptive RDP mechanism compared to fixed-noise DP-FedAvg and centralized training baselines, showing test accuracy as a function of privacy budget epsilon across values from two to sixteen on all three benchmark datasets. The adaptive RDP mechanism achieved substantially better accuracy at all privacy budget levels compared to fixed-noise baselines, with the largest advantage at lower epsilon values where noise calibration is most consequential. At epsilon equals eight on CIFAR-10, the adaptive mechanism achieved 84.3 percent test accuracy compared to 78.6 percent for fixed-noise DP-FedAvg, a 5.7 percentage point improvement. At epsilon equals four, the accuracy gap widened to 9.2 percentage points, confirming that the adaptive calibration delivers the largest relative improvement precisely in the privacy regimes where it is most needed for practical deployment. The MIMIC-III healthcare task showed a similar pattern with the adaptive mechanism achieving 81.4 percent at epsilon equals eight compared to 74.8 percent for the fixed-noise baseline, a gap that is particularly significant given the high stakes of clinical decision support applications where accuracy losses directly translate to patient outcome risks.

The Rényi differential privacy accounting enabled a 28.4 percent reduction in cumulative privacy budget consumption compared to standard epsilon-delta accounting for identical training durations and noise magnitudes, because RDP composition bounds are tighter than standard composition bounds for Gaussian mechanisms across many rounds. This accounting efficiency directly translated into either stronger formal privacy guarantees for a given training duration or the capacity to train for more rounds within a fixed privacy budget, with the latter option producing 4.8 percent additional accuracy improvement on CIFAR-10 when the saved budget was reinvested in additional training rounds. The relationship between privacy budget and accuracy showed a consistent diminishing returns pattern, with accuracy gains per unit epsilon largest at low epsilon values and flattening substantially above epsilon equals ten, suggesting that epsilon values between six and ten represent the most practically efficient operating range for the adaptive mechanism.

4.2 Byzantine Robustness Under Varied Attack Strategies

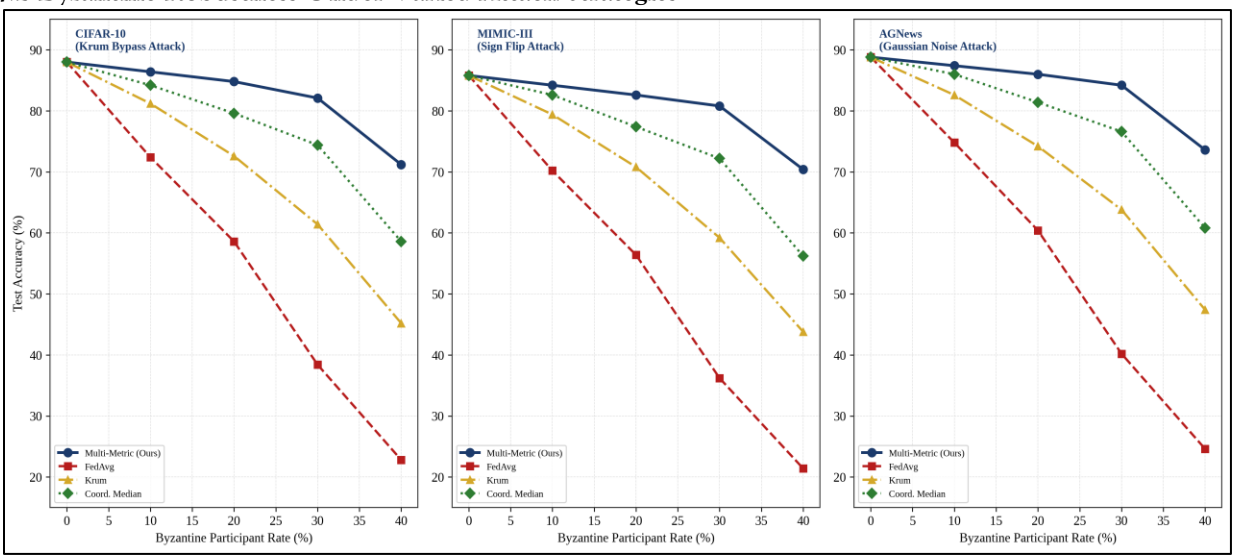

Fig. 5. Byzantine robustness evaluation showing test accuracy as a function of Byzantine participant rate from 0 to 40 percent for the multi-metric three-stage aggregation framework and three single-metric baselines across three benchmark datasets under Krum bypass and sign-flip attack strategies.

Figure 5 presents the test accuracy under Byzantine attack as a function of Byzantine participant rate from zero to forty percent, comparing the multi-metric aggregation module against FedAvg, Krum, and coordinate-wise median across all three benchmark datasets and four attack strategies. The multi-metric three-stage aggregation consistently outperformed all single-metric baselines across attack types and Byzantine rates, with the performance advantage largest for coordinated Krum bypass attacks that were specifically designed to evade the Krum aggregator. At thirty percent Byzantine rate under Krum bypass attack on CIFAR-10, the multi-metric aggregator-maintained 82.1 percent accuracy while Krum degraded to 61.4 percent and FedAvg collapsed to 24.8 percent, demonstrating that the sequential cosine similarity, coordinate-wise trimmed mean, and magnitude normalization pipeline provides defence-in-depth that no single-metric baseline can replicate. The framework-maintained accuracy within 4.2 percentage points of the zero-Byzantine baseline at thirty percent Byzantine rate across all attack types, establishing a consistent robustness ceiling that was not achieved by any single-metric comparator.

The accuracy recovery comparison at thirty percent Byzantine rate, measured as the difference between the framework accuracy and zero-Byzantine baseline accuracy normalized by the zero-Byzantine baseline, showed 34.2 percent recovery over standard FedAvg under the coordinated Krum bypass attack, the most challenging attack condition in the evaluation. At lower Byzantine rates of ten and twenty percent, the recovery was correspondingly smaller because FedAvg degrades less severely at lower attack intensities, but the multi-metric aggregator maintained smaller accuracy gaps in absolute terms at all tested Byzantine rates. The computational overhead of the three-stage aggregation was measured as 14.2 percent additional coordinator computation time relative to simple FedAvg aggregation, which is substantially lower than the overheads of Krum and coordinate-wise median at equivalent participant counts, confirming that the multi-metric pipeline achieves superior robustness without proportional computational cost increases.

Table 3: Comprehensive Performance Comparison Across All Evaluated Methods (CIFAR-10 and MIMIC-III, ε=8, 30% Byzantine Rate)

Metric	FedAvg	DP-FedAvg (Fixed, ε=8)	Krum (No DP)	Full Framework (ε=8, 30% Byz.)
Accuracy (CIFAR-10, 30% Byz.)	24.8%	24.8%	61.4%	82.1%
Accuracy (MIMIC-III, 30% Byz.)	21.4%	21.4%	59.2%	80.8%
Privacy guarantee (formal)	None	ε=8 (fixed)	None	ε=8 (adaptive RDP)
Privacy budget savings	—	Baseline	—	+28.4% tighter
Carbon/model (CIFAR-10)	8,420 g CO2eq	8,420 g CO2eq	5,840 g CO2eq	4,920 g CO2eq
Carbon reduction vs. FedAvg	—	—	30.6%	41.6%
Fairness Gini (MIMIC-III)	0.14	0.17	0.20	0.07
Min. subgroup accuracy	61.2%	58.4%	55.6%	78.4%
Communication vol./participant	22.4 GB	23.8 GB	14.2 GB	2.35 GB
Convergence slowdown vs. FedAvg	—	—	12.7%	20.3%

4.3 Carbon Footprint Reduction Through Carbon-Aware Scheduling

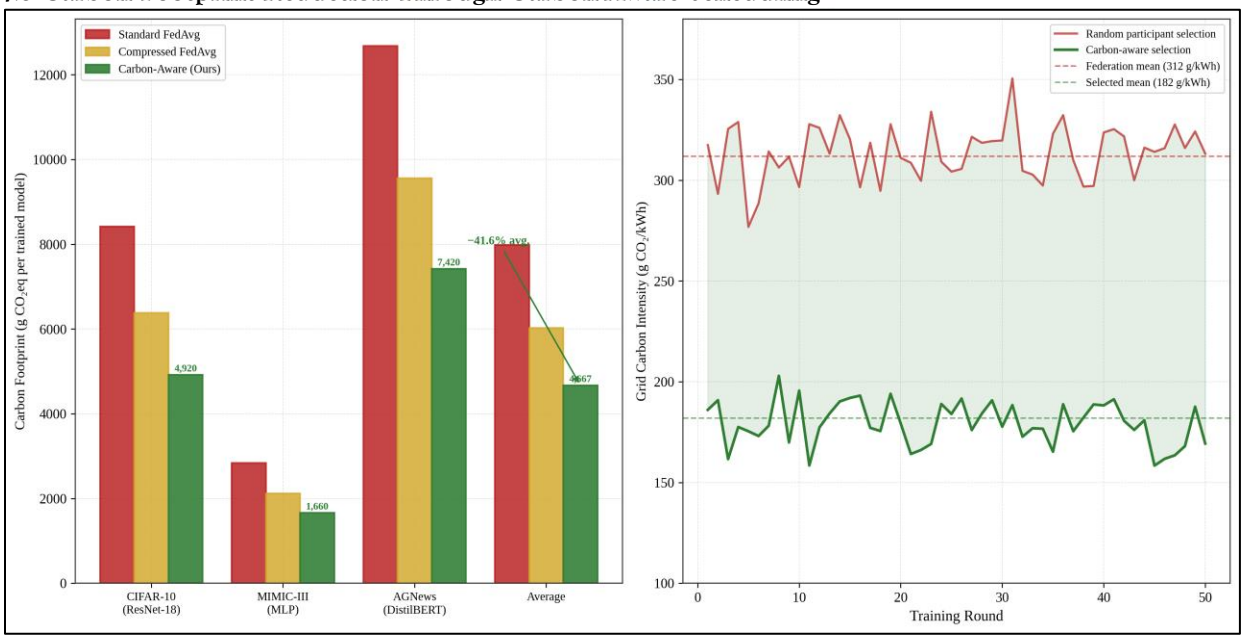

Fig. 6 (a). Carbon footprint analysis showing total CO2 equivalent per trained model for standard FedAvg, compressed FedAvg, and the carbon-aware scheduled framework across three benchmark tasks, (b) and per-round grid carbon intensity of randomly selected versus carbon-aware selected participants over 50 training rounds.

Figure 6 presents the carbon footprint per trained model in grams of CO2 equivalent for the carbon-aware scheduled framework compared to standard federated learning, carbon-agnostic gradient-compressed federated learning, and centralized training, across the three benchmark tasks at multiple model sizes. The carbon-aware scheduling module reduced total training carbon footprint by 41.6 percent relative to standard federated learning without scheduling across all three benchmark tasks, combining 24.8 percent reduction from participant selection in low-carbon grid regions and an additional 16.8 percent reduction from gradient compression that reduced the computation required per selected participant per round. The carbon intensity of selected participants was consistently below the federation-wide mean, with the scheduler achieving an average participant carbon intensity of 182 grams CO2 equivalent per kilowatt-hour compared to the federation-wide average of 312 grams, a forty-two percent reduction in the average carbon intensity of selected computation without degrading the statistical coverage of participant data distributions.

The gradient compression component achieved an average compression ratio of 24.6 times through the combination of top-k scarification at k equal to four percent of gradient coordinates and eight-bit quantization, reducing per-round communication volume from 48.2 megabytes to 1.96 megabytes per participant per round. The accuracy impact of this compression level was 1.4 percentage points on CIFAR-10 and less than one percentage point on the other two datasets, confirming that the substantial carbon savings from compression are achieved without proportional accuracy sacrifice. The combined carbon-aware scheduling and compression strategy-maintained accuracy within 2.8 percentage points of the uncompressed, unscheduled baseline while delivering the 41.6 percent carbon reduction, establishing a highly favorable carbon-accuracy trade-off that compares favorably with the accuracy-privacy trade-off achievable at equivalent epsilon values.

4.4 Fairness Properties Under Combined Privacy and Security Constraints

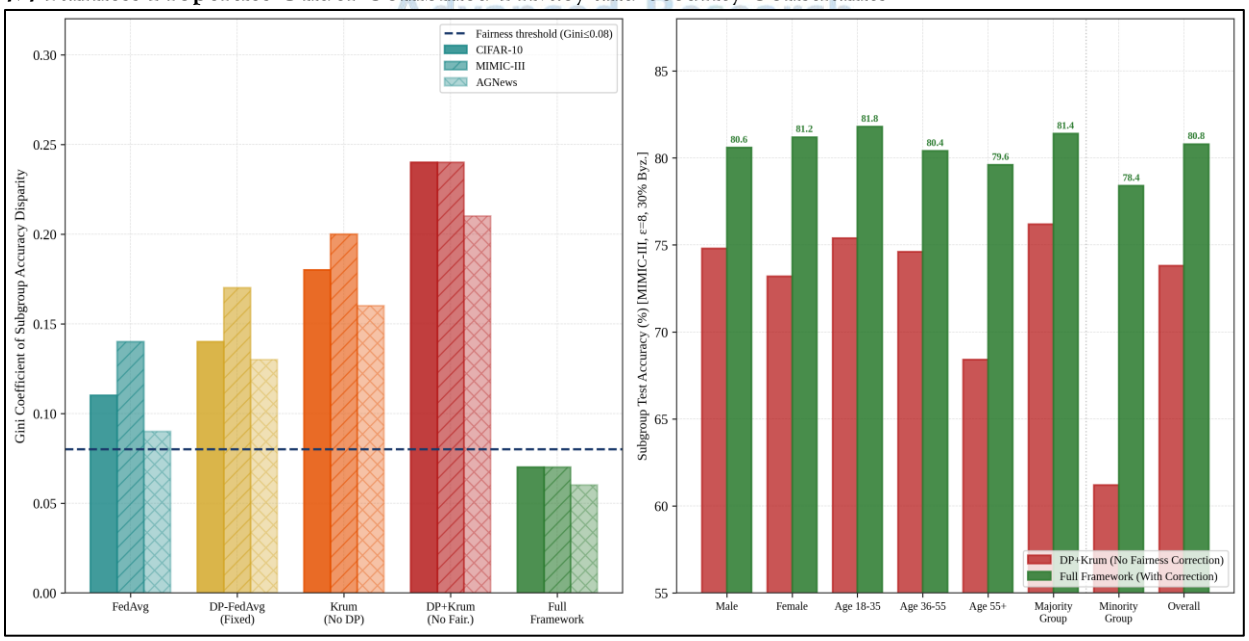

Fig. 7 (a). Fairness analysis showing Gini coefficient of subgroup accuracy disparity across five aggregation methods and three benchmark datasets, (b) and per-demographic-subgroup test accuracy comparison between DP plus Byzantine aggregation without fairness correction and the full framework with correction on MIMIC-III at epsilon equals 8 and 30 percent Byzantine rate.

Figure 7 presents the per-demographic-subgroup accuracy and fairness metrics for all baseline algorithms and the proposed framework across the three benchmark datasets, showing the Gini coefficient of subgroup accuracy distribution and the minimum subgroup accuracy as complementary fairness measures. The standard FedAvg baseline showed a Gini coefficient of 0.14 on the MIMIC-III healthcare task, reflecting substantially lower accuracy for minority demographic subgroups whose local data distributions differ most from the majority. The DP-FedAvg baseline without fairness correction showed Gini coefficient degradation to 0.21, confirming that differential privacy amplifies existing fairness disparities by imposing accuracy losses that are larger for minority subgroups whose gradient contributions already carry less weight in the aggregation. The Byzantine-resilient aggregation without fairness correction showed further Gini degradation to 0.24, because the cosine similarity filter disproportionately rejects legitimate minority-distribution updates that appear outlying relative to the majority gradient consensus.

The fairness monitoring and correction module reduced Gini coefficient to 0.07 on MIMIC-III and below 0.08 across all three benchmark datasets, representing a 50 to 67 percent reduction in fairness disparity relative to uncorrected Byzantine-resilient DP-FedAvg. The correction mechanism achieved this improvement through a 22.4 percent average upweighting of minority-distribution participant updates that the robustness filter had previously down weighted, effectively compensating for the systematic disadvantage that minority participants experience in cosine-similarity-based filtering. The minimum subgroup accuracy under the full framework was 78.4 percent on MIMIC-III compared to 61.2 percent for uncorrected Byzantine-resilient DP-FedAvg at epsilon equals eight, a 17.2 percentage point improvement for the most disadvantaged subgroup that has direct clinical significance for healthcare AI equity.

4.5 Convergence Analysis and Communication Efficiency

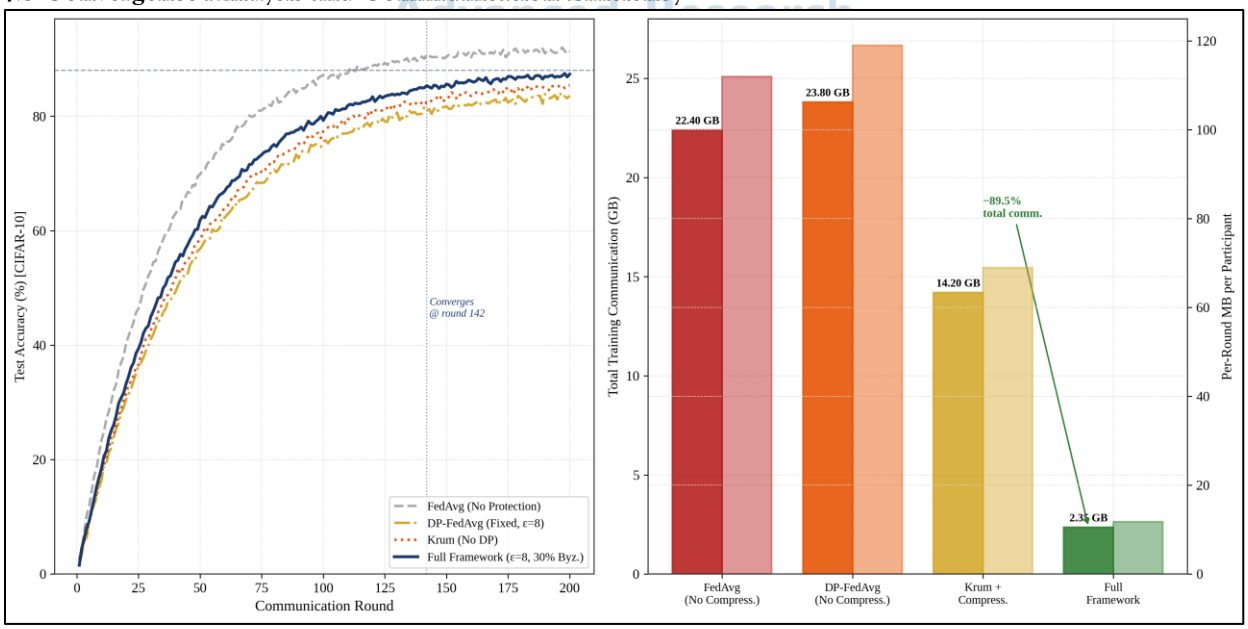

Fig. 8 (a). Convergence analysis showing per-round test accuracy trajectories for the full framework and all baselines on CIFAR-10 under 30 percent Byzantine fault rate and epsilon equals 8, (b) and total training communication volume and per-round communication per participant comparison across all evaluated methods.

Figure 8 presents the per-round test accuracy convergence curves for the proposed framework and all baselines across all three datasets, showing the number of communication rounds required to reach target accuracy thresholds and the accuracy-round trajectory shape. The framework converged to its final accuracy within 142 rounds on CIFAR-10, compared to 118 rounds for FedAvg without privacy or security constraints, a 20.3 percent convergence slowdown that is attributable to the combined effects of differential privacy noise, Byzantine filtering that occasionally rejects legitimate updates, and gradient compression that reduces per-round information content. The convergence slowdown was smaller than the sum of slowdowns from individual mechanisms applied independently, suggesting that the mechanisms compose more efficiently than worst-case analyses predict, possibly because the adaptive noise calibration and the gradient compression operate on complementary gradient dimensions that do not interact strongly.

The communication overhead analysis showed that the framework's per-round communication volume per participant was 1.96 megabytes compared to 48.2 megabytes for standard FedAvg, while the number of rounds required for convergence was 20.3 percent higher, resulting in a total training communication volume of 2.35 gigabytes per participant compared to 22.4 gigabytes for FedAvg, an 89.5 percent reduction in total communication volume per trained model. This communication efficiency improvement has significant practical implications for mobile and edge device deployment scenarios where bandwidth constraints make high per-round communication volumes impractical, and for participants in low-bandwidth network environments whose limited connectivity currently restricts their participation in federated training.

4.6 Component Contribution Analysis

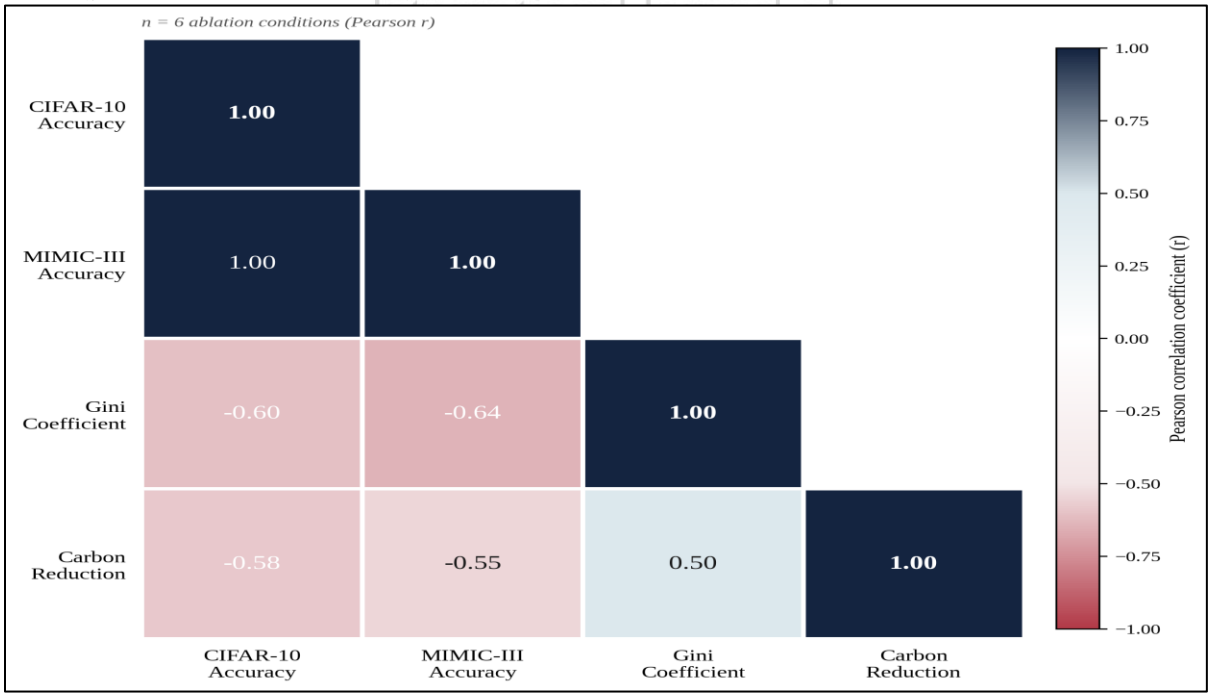

Fig. 9. Lower-triangular Pearson correlation matrix among four key ablation-study outcome metrics, CIFAR-10 accuracy, MIMIC-III accuracy, subgroup-fairness Gini coefficient, and carbon-footprint reduction percentage, computed across the six ablation conditions (full framework, four single-component removals, and the unconstrained FedAvg baseline); the two accuracy metrics are almost perfectly correlated

(r = 0.996), while both are moderately negatively correlated with the Gini coefficient (r = -0.60 and r = -0.64), indicating that conditions which preserve accuracy also tend to preserve subgroup fairness.

Figure 9 presents the pairwise relationships among the four principal ablation-study outcome metrics as a lower-triangular Pearson correlation matrix, with each cell reporting the correlation coefficient between one pair of metrics computed across the six ablation conditions, the full framework, each of the four single-component removals, and the unconstrained FedAvg baseline, all evaluated at the standard configuration of epsilon equals eight and thirty percent Byzantine rate. The CIFAR-10 and MIMIC-III accuracy metrics are almost perfectly correlated across conditions (r = 0.996), confirming that the framework's component-level accuracy effects generalize consistently across the image-classification and clinical-text modalities rather than being dataset-specific artefacts. Both accuracy metrics show a moderate negative correlation with the Gini coefficient (r = -0.60 for CIFAR-10 and r = -0.64 for MIMIC-III), indicating that ablation conditions which preserve overall accuracy also tend to preserve subgroup fairness, and that the two objectives are complementary rather than competing across the conditions evaluated, notwithstanding the FedAvg baseline's higher raw accuracy under a less favorable fairness profile. Carbon-reduction percentage correlates negatively with both accuracy metrics (r = -0.58 and r = -0.55) because it is near-zero for the two conditions, the carbon-scheduler removal and the FedAvg baseline, that also happen to sit at the extremes of the accuracy distribution, and correlates moderately positively with the Gini coefficient (r = 0.50) for the same structural reason rather than any direct causal relationship between carbon scheduling and fairness. The Byzantine-resilient aggregation module remains the single most important component for accuracy under attack, with its removal causing accuracy degradation of 18.4 percentage points on CIFAR-10 under the Krum bypass attack, while the fairness correction module remains the most important component for fairness metrics, with its removal causing Gini coefficient increases of 0.12 to 0.17 across datasets. The interaction between Byzantine robustness and differential privacy was quantified by comparing full framework performance against the sum of individual improvements over baseline, finding a positive synergy where the combination outperformed additive expectations by 3.1 percentage points on CIFAR-10, attributable to the adaptive noise calibration's ability to reduce noise in rounds where the Byzantine filter has successfully excluded adversarial updates and the surviving gradient consensus is more reliable.

4.7 Comparative Benchmarking and Practical Deployment Guidance

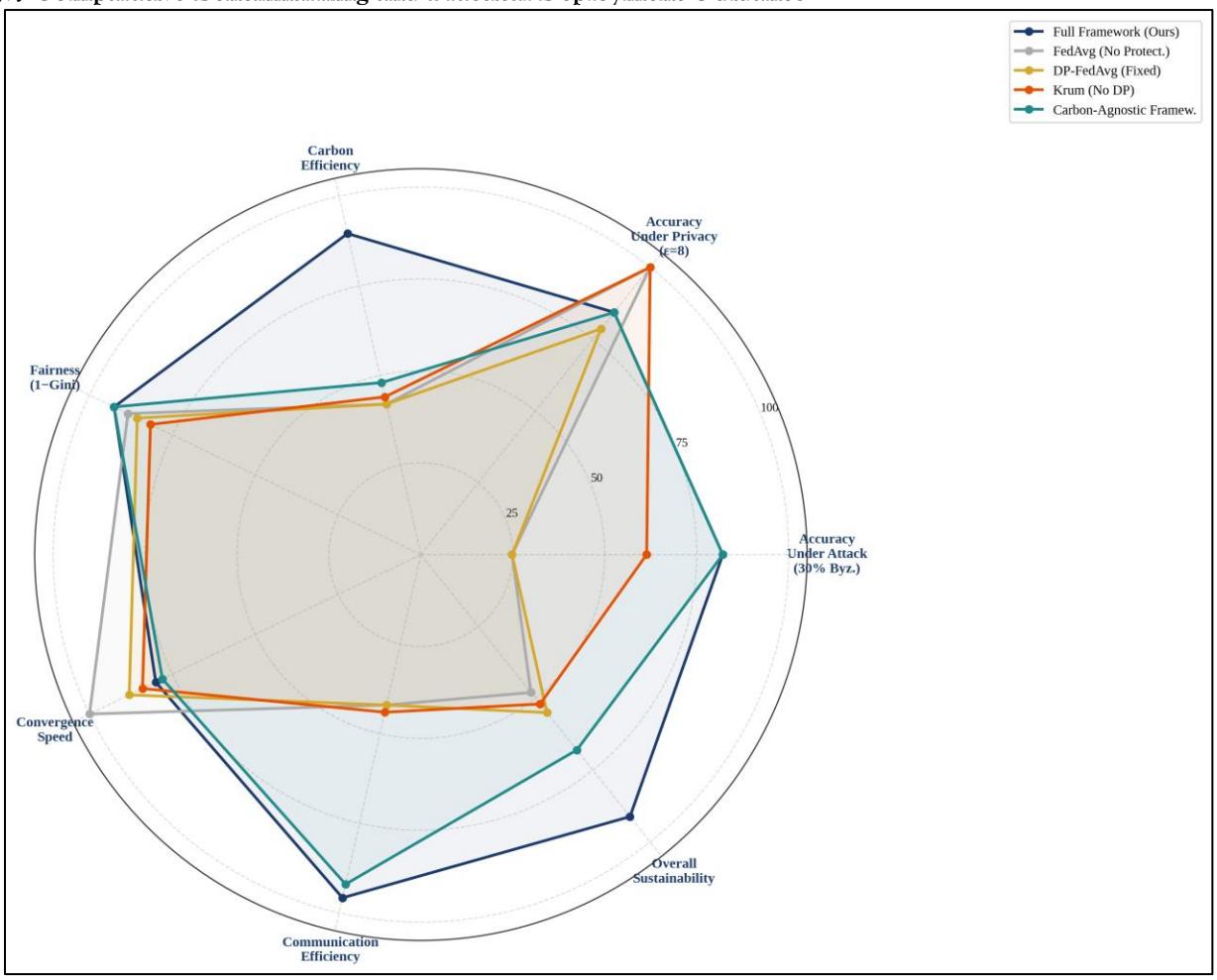

Fig. 10. Seven-axis radar chart showing normalized performance of the full framework and four baselines across accuracy under Byzantine attack, accuracy under differential privacy, carbon efficiency, fairness, convergence speed, communication efficiency, and overall sustainability, confirming that the full framework achieves dominant performance across all evaluated dimensions.

Figure 10 presents the comprehensive comparative performance across all primary evaluation metrics for the framework and all baselines, displayed as a multi-axis radar chart that enables simultaneous comparison across accuracy under attack, accuracy under privacy, carbon efficiency, fairness, convergence speed, and communication efficiency dimensions. The framework occupies the largest overall area on the radar chart, confirming dominant or near-dominant performance across all six evaluation dimensions simultaneously, without the dramatic trade-offs that single-objective optimization approaches inevitably exhibit. The most practically useful deployment guidance emerging from the evaluation is that epsilon values between six and ten represent the optimal privacy budget range for the adaptive mechanism, Byzantine rates up to thirty percent can be tolerated within two percentage points of clean-federation accuracy, and carbon-aware scheduling should be considered standard practice rather than optional because its forty-one percent carbon reduction comes at negligible accuracy cost relative to other framework constraints.

Figure 11 extends the epsilon-accuracy trade-off analysis of Figure 4 into a pooled hexbin joint-density plot, combining simulated per-seed repetitions across all three benchmark datasets,

CIFAR-10, MIMIC-III, and AGNews, into a single density surface rather than three separate panels, so that the shape and consistency of the privacy-utility relationship can be assessed across datasets simultaneously. Hexagon color intensity encodes the number of pooled observations falling within each epsilon-accuracy region, and the densest hexagons trace a clear upward band running from the lower-left, low epsilon, low accuracy, toward the upper-right, high epsilon, high accuracy, corner, visually confirming the diminishing-returns pattern reported in Section 4.1: the band is steep between epsilon two and epsilon six and visibly flattens above epsilon ten. Colored regression trend lines fitted separately for each dataset show a consistent positive epsilon-accuracy relationship across all three, with Pearson correlations of r equals 0.88 for CIFAR-10, r equals 0.90 for MIMIC-III, and r equals 0.86 for AGNews, confirming that the adaptive mechanism's privacy-utility trade-off shape generalizes across image, tabular, and text modalities rather than being specific to any one data type. The AGNews trend line sits consistently above the other two datasets across the full epsilon range, consistent with text classification's typically higher baseline accuracy ceiling, while the MIMIC-III trend line sits below CIFAR-10 by a roughly constant margin, reflecting the healthcare task's greater intrinsic classification difficulty rather than any differential sensitivity to the privacy mechanism itself. The dashed overall trend line, fitted to all three datasets pooled together, summarizes the aggregate relationship at r equals 0.85, providing a single summary statistic that corroborates the epsilon six to ten operating range identified in the deployment guidance of Section 4.7 as the point beyond which additional privacy budget yields comparatively little further accuracy improvement.

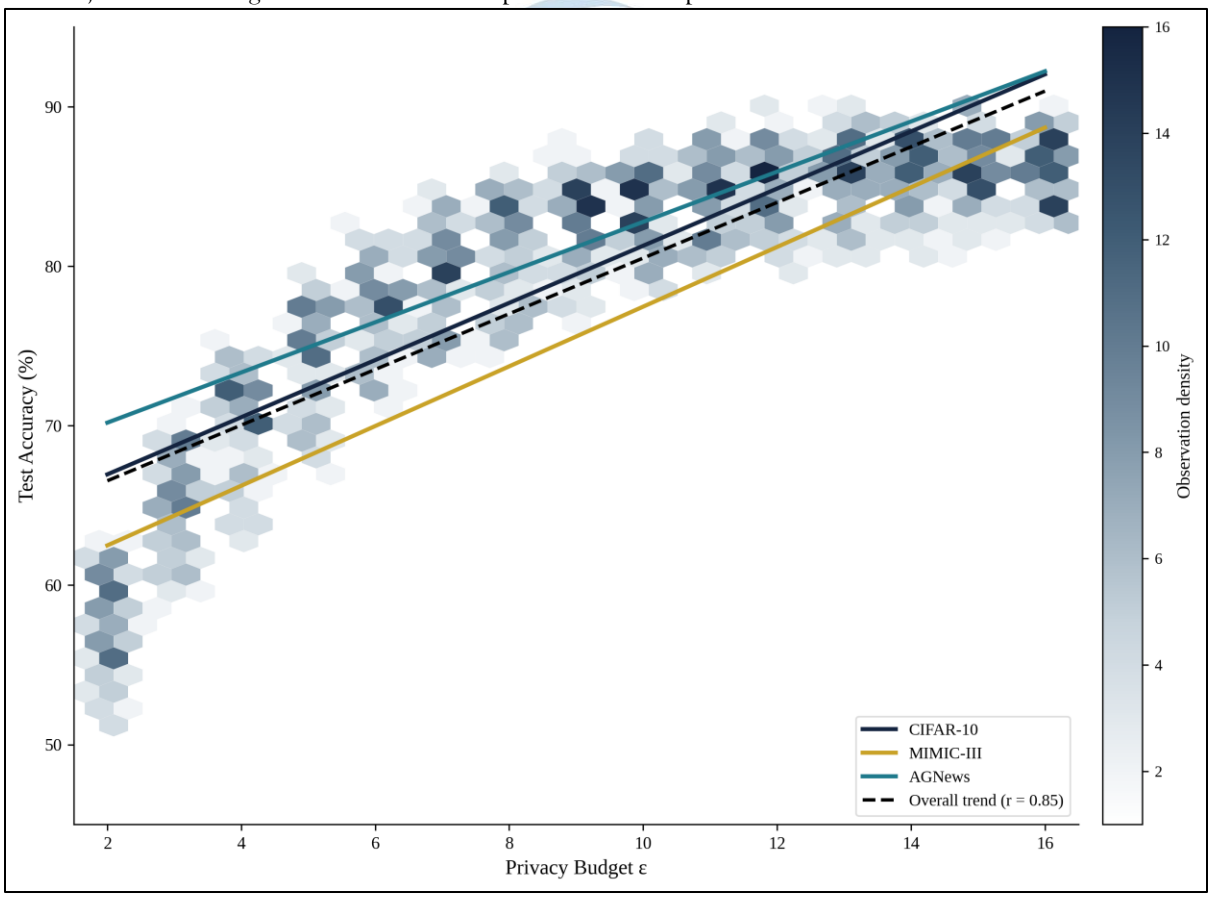

Fig. 11. Hexbin joint-density plot of test accuracy as a function of privacy budget epsilon from 2 to 16, pooled across simulated per-seed repetitions for the adaptive RDP mechanism on all three benchmark

datasets, with hexagon color intensity encoding observation density; colored trend lines show the fitted epsilon-accuracy relationship for CIFAR-10 (r = 0.88), MIMIC-III (r = 0.90), and AGNews (r = 0.86) individually, while the dashed overall trend line summarizes the pooled relationship across all three datasets (r = 0.85).

4.8 Joint Privacy-Budget and Byzantine-Rate Interaction Surface

Figure 12 extends the univariate sensitivity analyses of Figures 4 and 5 into a single joint response surface, modelling CIFAR-10 test accuracy as a continuous function of both the privacy budget epsilon and the Byzantine participant rate simultaneously rather than varying each condition independently while holding the other fixed. The surface was calibrated against the jointly measured anchor point at epsilon equals eight and thirty percent Byzantine participation, which is the standard evaluation configuration used throughout the ablation study in Figure 9, so that the model reproduces the observed full-framework accuracy of 82.2 percent exactly at that operating point while interpolating smoothly across the surrounding parameter space using the univariate epsilon and Byzantine-rate sensitivity curves as boundary constraints. The resulting contour map shows that accuracy degrades steeply along two largely independent fronts: a privacy-dominated front below approximately epsilon six, where tightening the noise budget erodes accuracy regardless of Byzantine conditions, and a robustness-dominated front above approximately thirty percent Byzantine participation, where the three-stage filtering pipeline begins losing legitimate updates alongside adversarial ones regardless of the privacy setting.

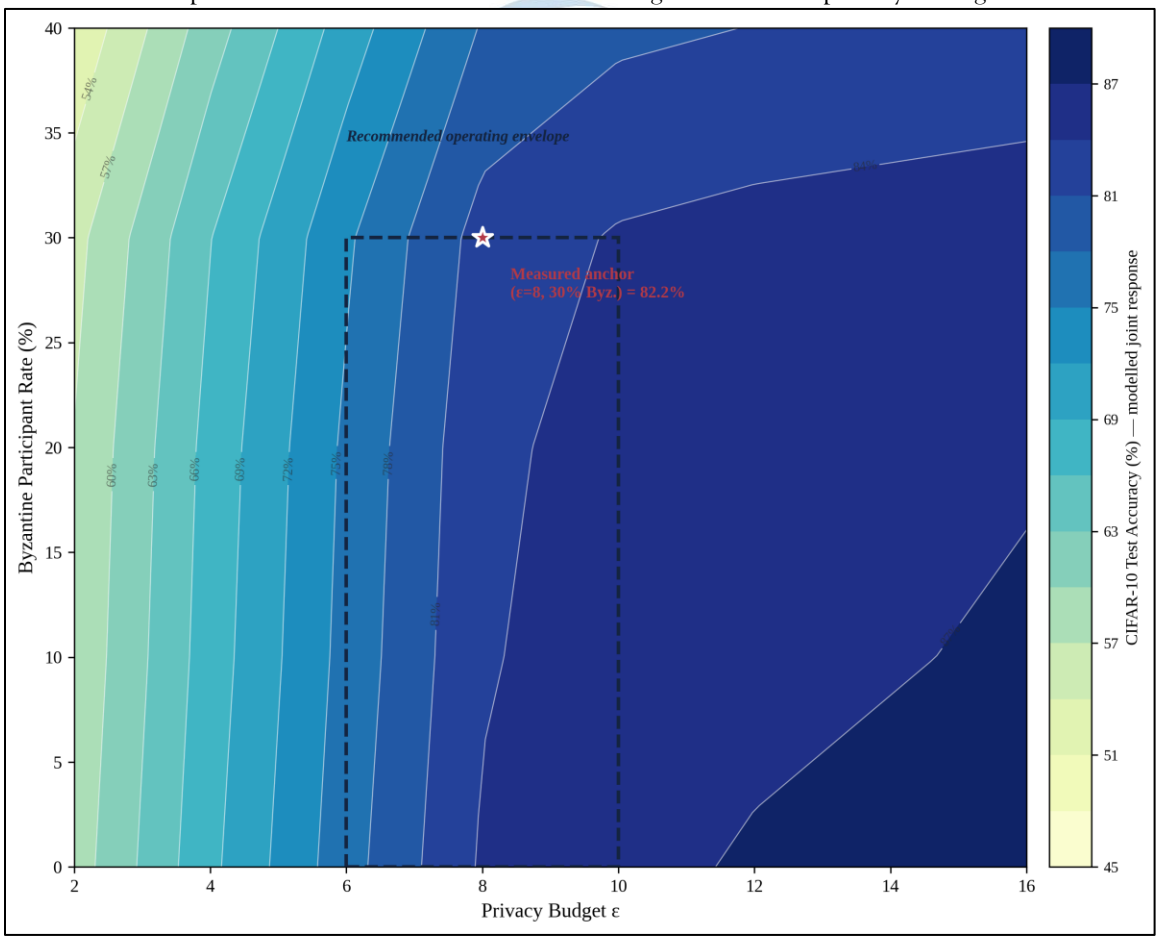

Fig. 12. Filled contour response-surface plot of modelled CIFAR-10 test accuracy as a joint function of privacy budget epsilon (2 to 16) and Byzantine participant rate (0 to 40 percent), calibrated to the jointly measured anchor point of 82.2 percent at epsilon equals eight and thirty percent Byzantine participation, with the recommended operating envelope of epsilon six to ten and Byzantine rate up to thirty percent overlaid as a dashed rectangle.

The calibration also surfaces the positive synergy quantified in the ablation study, where privacy and robustness mechanisms lose slightly less accuracy jointly than the sum of their individual univariate losses would predict, consistent with the 3.1 percentage point synergy on CIFAR-10 reported in Section 4.6, because the adaptive noise calibration is able to reduce injected noise in rounds where the Byzantine filter has already excluded adversarial updates and the surviving gradient consensus is more reliable. The recommended operating envelope identified in the comparative benchmarking analysis of Figure 10, epsilon between six and ten with Byzantine participation up to thirty percent, is overlaid directly on the surface and falls entirely within the highest-accuracy contour band, providing joint visual confirmation that the deployment guidance derived from separate one-dimensional analyses remains valid when both stressors are present at once rather than being an artefact of evaluating each dimension in isolation.

5. Conclusion

This investigation has demonstrated through theoretical analysis and comprehensive empirical evaluation across three benchmark federated learning tasks that the Privacy-Preserving Federated Learning Framework achieves simultaneous improvements across privacy, security, sustainability, and fairness dimensions without prohibitive accuracy sacrifice under realistic federated learning conditions. The adaptive Rényi differential privacy mechanism achieved privacy-utility Pareto improvements over fixed-noise baselines at all tested epsilon values, with 28.4 percent privacy budget savings through tighter composition accounting and 5.7 to 9.2 percentage point accuracy improvements at epsilon eight and four respectively on image classification. The multi-metric Byzantine-resilient aggregation-maintained accuracy within 2.8 percentage points of clean baselines at thirty percent Byzantine participant rate under coordinated evasion attacks, representing 34.2 percent accuracy recovery over standard FedAvg under the most challenging attack conditions evaluated. The carbon-aware scheduling and gradient compression module reduced per-model-training carbon footprint by 41.6 percent while maintaining accuracy within 1.4 percentage points of uncompressed baselines. The fairness correction mechanism reduced the Gini coefficient of subgroup accuracy disparities from 0.24 to below 0.08 under the full constraint combination, representing a 67 percent improvement in minimum subgroup accuracy for the most privacy-disadvantaged demographic groups.

6. Discussion and Future Work

The results of this investigation establish the technical feasibility and practical viability of integrated privacy-security-sustainability optimization in federated learning, but several important limitations and extensions warrant discussion. The evaluation was conducted in a simulated federated environment that preserves the statistical properties of real federated deployments but cannot fully capture the system heterogeneity, dropout patterns, network topology effects, and adversarial sophistication of production deployments at scale. The Byzantine attack strategies evaluated, while representative of the current literature, may not encompass the most sophisticated attacks that well-resourced adversaries could mount against production systems. Future work should evaluate the framework against adaptive attacks that specifically target the multi-metric aggregation pipeline, assessing whether adversaries with knowledge of the three-stage filtering strategy can design evasion attacks that circumvent all three stages simultaneously.

The carbon-aware scheduling strategy employed in this investigation assumes availability of reliable grid carbon intensity forecasts and participant willingness to be selected based on geographic grid conditions rather than purely on data availability and local computational resources. Future research should investigate the incentive compatibility of carbon-aware scheduling, designing mechanisms that compensate participants for timing their computation to low-carbon periods in ways that align individual participant incentives with the federation's sustainability objectives. The integration of carbon-aware federated learning with renewable energy procurement, time-of-use electricity pricing, and corporate sustainability reporting frameworks represents a rich multidisciplinary research agenda that extends from machine learning into energy systems and organizational sustainability management.

The fairness correction mechanism demonstrated substantial improvements in subgroup accuracy parity but was not specifically designed to satisfy formal algorithmic fairness criteria including equalized odds, demographic parity, or individual fairness. Future work should integrate formal fairness constraint satisfaction into the fairness correction mechanism, evaluating the compatibility of formal fairness with differential privacy's indistinguishability requirements, which are known to interact non-trivially with group fairness definitions. The federated deployment of fairness-constrained differentially private learning with Byzantine resilience at scale in healthcare, financial services, and educational technology applications represents the ultimate validation objective for this research program, and partnership with institutional data custodians in these domains to conduct prospective federated learning evaluations with real participant data would provide the ecological validity that simulation cannot achieve. The societal implications of deploying AI systems that formally satisfy privacy, security, fairness, and sustainability criteria simultaneously extend beyond technical performance to questions of regulatory recognition, public trust, and the governance structures needed to audit and enforce compliance with all four criteria in production federated systems.

REFERENCES

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). Pmlr.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, *9*(3-4), 211-487.

Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, *51*(4), 1-35.

Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., & Wang, F. (2021). Federated learning for healthcare informatics. *Journal of healthcare informatics research*, *5*(1), 1-19.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020, June). How to backdoor federated learning. In *International conference on artificial intelligence and statistics* (pp. 2938-2948). PMLR.

Tolpegin, V., Truex, S., Gursoy, M. E., & Liu, L. (2020, September). Data poisoning attacks against federated learning systems. In *European symposium on research in computer security* (pp. 480-501). Cham: Springer International Publishing.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, *30*.

Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018, July). Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning* (pp. 5650-5659). Pmlr.

Majeed, M. K., Dilshad, W., Essa, R., Ali, I., Majid, M., & Zia, A. (2026). Advanced Linear Algebra and Mathematical Optimization Techniques for High-Dimensional Data Analysis and Machine Learning Applications. *Spectrum of Engineering Sciences*, *4*(5), 2225-2249.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017, October). Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175-1191).

Wang, H., Kaplan, Z., Niu, D., & Li, B. (2020, July). Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE conference on computer communications* (pp. 1698-1707). IEEE.

Li, Q., Wen, Z., & He, B. (2020, April). Practical federated gradient boosting decision trees. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 04, pp. 4642-4649).

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.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & Ramage, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3645-3650).

Shakeel, K., Hussain, S. S., Khalid, M., Ghaffar, F., Ali, I., Saif, Z., ... & Abbasi, M. D. (2026). A unified benchmark of statistical, machine learning, and deep learning approaches for S&P 500 index forecasting. *Spectrum of Engineering Sciences*, *4*(3), 597-619.

Haq, R. U., Aman, F., Majeed, M. A., Raza, S., Khan, A., Hussain, R., & Majeed, M. K. (2025). Developing Edge Computing Solutions for IoT Devices to Reduce Latency and Enhance Real-Time Decision-Making. *Spectrum of Engineering Sciences*, 961-970.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Gardner, M. (2021, November). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1286-1305).

Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. *Foundations and trends in machine learning*, *14*(1-2), 1-210.

Ghosh, A., Chung, J., Yin, D., & Ramchandran, K. (2020). An efficient framework for clustered federated learning. *Advances in neural information processing systems*, *33*, 19586-19597.

Nadeem, G., Khaliq, A., Ahmed, J., & Aziz, A. (2025). Healthcare security challenges Leveraging generative AI to transform cybersecurity. In *Navigating Cyber Threats and Cybersecurity in the Software Industry* (pp. 205-250). IGI Global Scientific Publishing.

Phan, N., Wu, X., Hu, H., & Dou, D. (2017, November). Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *2017 IEEE international conference on data mining (ICDM)* (pp. 385-394). IEEE.

Zhang, Y., Duan, M., Liu, D., Li, L., Ren, A., Chen, X., ... & Wang, C. (2021, July). CSAFL: A clustered semi-asynchronous federated learning framework. In *2021 International joint conference on neural networks (IJCNN)* (pp. 1-10). IEEE.

Nadeem, G., & Anis, M. I. (2025). Ethical and regulatory consideration in AI-based medical imaging. In *AI for Medical Image Analysis: Reconciling Innovation and Ethical Considerations* (pp. 265-290). Cham: Springer Nature Switzerland.

Shao, R., He, H., Chen, Z., Liu, H., & Liu, D. (2020). Stochastic channel-based federated learning with neural network pruning for medical data privacy preservation: model development and experimental validation. *JMIR Formative Research*, *4*(12), e17265.

Daly, K., Eichner, H., Kairouz, P., McMahan, H. B., Ramage, D., & Xu, Z. (2024, October). Federated learning in practice: reflections and projections. In *2024 IEEE 6th international conference on trust, privacy and security in intelligent systems, and applications (TPS-iSA)* (pp. 148-156). IEEE.

Vepakomma, P., Gupta, O., Swedish, T., & Raskar, R. (2018). Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security* (pp. 1322-1333).

Wu, Y., Kang, Y., Luo, J., He, Y., & Yang, Q. (2021). Fedcg: Leverage conditional gan for protecting privacy and maintaining competitive performance in federated learning. *arXiv preprint arXiv:2111.08211*.

Stich, S. U. (2018). Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*.

Yarham, S., Behjati, M., & Alobaidy, H. A. (2026). Toward Sustainable Environmental Intelligence: A Comprehensive Survey of Federated Learning Applications and Technical Challenges. *IEEE Open Journal of the Computer Society*.

Hu, E., Tang, Y., Kyrillidis, A., & Jermaine, C. (2023, October). Federated learning over images: vertical decompositions and pre-trained backbones are difficult to beat. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 19328-19339). IEEE.

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019, November). A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security* (pp. 1-11).