Defending Federated Learning: Adaptive Integration of Differential Privacy, SMPC, and Byzantine Robustness

Syed Momin Naqvi [1] **Faisal Iradat** [1]

[1] *Department of Computer Science, School of Mathematics and Computer Science, Institute of Business Administration (IBA), Karachi, Pakistan*. Correspondence to: Syed Momin Naqvi momin.naqvi.31515@khi.iba.edu.pk, Faisal Iradat firadat@iba.edu.pk.

1. Introduction

Federated Learning (FL) allows multiple parties to jointly train models without centrally storing raw data. It has become a revolutionary paradigm for collaborative ML [1]. Since HIPAA and GDPR forbid data sharing in certain industries such as healthcare, finance, and collaborative cybersecurity, Federated Learning becomes the default.

However, standard FL implementations face three critical vulnerability classes. **First**, gradient information shared during training leaks sensitive details through inference attacks: model inversion can reconstruct training samples [2], while membership inference determines whether specific records were used in training [3]. **Second**, the distributed nature makes FL vulnerable to Byzantine adversaries who poison models through corrupted local data or malicious gradient manipulation [4, 5]. **Third**, transmitted updates lack cryptographic protection, exposing them to eavesdropping and server surveillance.

Existing solutions employ single-layered defenses. Differential Privacy (DP)-based approaches [6, 7] provide statistical privacy but no attack protection; Secure Multi-Party Computation (SMPC)-based frameworks [8] encrypt aggregation but lack DP guarantees; Byzantine-robust methods [9, 10] detect attacks but expose gradients. No existing framework systematically integrates these mechanisms or adapts to evolving threats.

This paper proposes the **Hybrid Federated Learning (HFL) Framework**, the first unified architecture integrating: (1) Differential Privacy for statistical indistinguishability, (2) Secure Multi-Party Computation for cryptographic confidentiality, (3) Robust Aggregation for Byzantine resilience, and (4) an adaptive control module dynamically balancing privacy budgets (ε), robustness thresholds (τ), and participation rates (ρ) based on real-time metrics.

2. Related Work and Gap Analysis

Table 1 compares representative FL frameworks across key security dimensions. Standard FedAvg [1] provides no explicit security guarantees beyond data localization. DP-FedAvg [6] and Local DP approaches [7] offer privacy through gradient perturbation but remain vulnerable to model poisoning. Secure Aggregation [8] uses SMPC to hide individual updates but lacks statistical privacy guarantees. Byzantine-robust aggregators including Krum [9], Trimmed Mean [10], and RFA [11] detect adversarial updates but expose gradient information. Recent hybrid approaches [12] combine DP with robust aggregation but treat them as independent modules, failing to address their interaction dynamics.

Table 1: Comparison of federated learning frameworks. Legend: ✓= Supported; × = Not supported.

Framework	DP	SMPC	Robust	Adaptive
FedAvg [1]	×	×	×	×
DP-FedAvg [6]	✓	×	×	×
Secure Agg [8]	×	✓	×	×
Krum [9]	×	×	✓	×
RFA [11]	✓	×	✓	×
Proposed HFL	✓	✓	✓	✓

The identified research gap is clear: *no existing framework provides unified, adaptive integration of complementary privacy-preserving and adversarial-resilient mechanisms while dynamically managing their inherent trade-offs*.

3. Proposed HFL Framework

3.1 Architecture Overview

The HFL framework operates through sequential layered defenses with continuous feedback control (Figure 1). Distributed clients perform local training on private datasets $D_1, D_2, \ldots, D_N$, computing gradient updates ∇_i that never leave their premises. These updates pass through integrated privacy mechanisms before secure aggregation and robust filtering produce the global model update.

Layer 1: Differential Privacy. Each client applies calibrated Gaussian noise to local gradients:

$$\widetilde{\nabla}_i = \nabla_i + \mathcal{N}(0, \sigma^2 I), \tag{1}$$

where σ is determined by the privacy budget ε to satisfy (ε, δ)-differential privacy:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta, \tag{2}$$

for adjacent datasets D and D' differing by one record.

Layer 2: Secure Multi-Party Computation. Noisy gradients are secret-shared across m non-colluding servers using additive sharing such that $\sum_{j=1}^{m} s_{i,j} = \widetilde{\nabla}_i$. Each server receives one share from each client, enabling computation of $\sum_{i=1}^{K} \widetilde{\nabla}_i$ without any single server accessing individual gradients [8].

Layer 3: Robust Aggregation. Before incorporating the aggregated update, robust filtering detects and

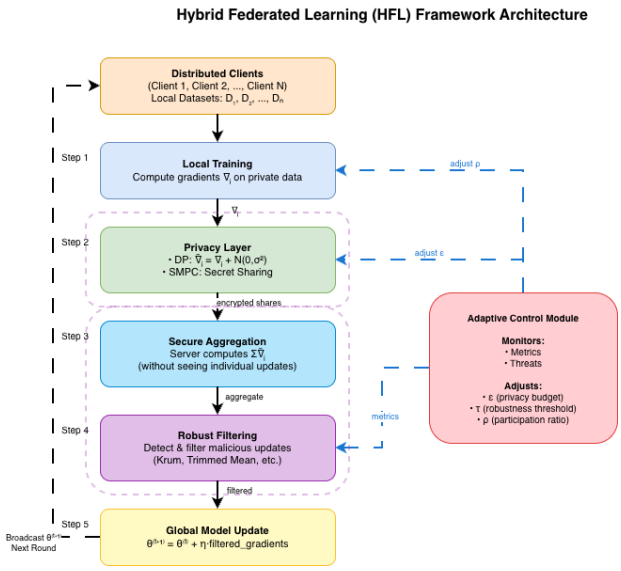

Fig. 1: Architecture of the Hybrid Federated Learning Framework showing layered defenses and adaptive control feedback loop.

suppresses Byzantine outliers. Krum [9], which selects updates from the dense cluster of honest participants, or Trimmed Mean [10] has been used. Updates exceeding anomaly threshold τ are excluded.

Adaptive Control Module. This component continuously monitors convergence rate $\Delta\mathcal{L}^{(t)}$, anomaly density (proportion of flagged updates), and resource utilization. It dynamically adjusts ε, τ, and ρ. When high anomaly density indicates attacks, the system decreases ε (more noise), tightens τ (stricter filtering), and reduces ρ (selective participation).

3.2 Theoretical Foundation

Secure federated learning as multi-objective optimization has been formulated as :

$$\min_{\varepsilon,\tau,\rho} \mathcal{L}(\varepsilon,\tau,\rho) = \alpha \cdot \mathcal{L}_{\text{acc}} + \beta \cdot \mathcal{L}_{\text{priv}} + \gamma \cdot \mathcal{L}_{\text{rob}}, \quad (3)$$

where $\mathcal{L}_{\text{acc}}$ quantifies model accuracy degradation, $\mathcal{L}_{\text{priv}}$ measures privacy leakage, and $\mathcal{L}_{\text{rob}}$ captures adversarial vulnerability. These domain-specific coefficients α, β, γ signify the application priorities (healthcare emphasizes β, cybersecurity emphasizes γ).

Key Trade-offs: By decreasing ε privacy can be enhanced, but it will eventually degrade accuracy and obscure hostile or adverse signals. If tune τ is tuned, it will improve robustness, but this risks rejecting legitimate noisy updates. By reducing ρ the computation can be lessened, but then limit the representativeness. The adaptive control module seeks Pareto-optimal configurations that balance these competing objectives.

Privacy-Robustness Conflict Resolution: A key challenge is the interference between DP noise and anomaly detection. When ε is low (high noise), legitimate updates can be flagged as anomalous. Our adaptive mechanism handles this by relaxing τ proportionally. This reduces false positives without sacrificing attack detection when noise levels are lower.

4. Key Contributions and Novelty

This framework makes five contributions to federated learning security:

- **Multi-layered integration**: Robust aggregation, SMPC, and DP have been used as unified defense layers rather than separate modules. To our knowledge, no existing framework combines all three systematically.
- **Adaptive parameter tuning**: The system adjusts ε, τ, and ρ in real time based on resource constraints, threat indicators, and convergence behavior.
- **Privacy-robustness conflict resolution**: Formalize and solve the problem where DP noise interferes with Byzantine detection, causing false positives.
- **Multi-objective optimization**: Equation 3 formalizes the three-way trade-off between accuracy, privacy, and robustness.
- **Threat coverage**: The framework protects against transmission exposure, Byzantine attacks, and inference attacks simultaneously.

5. Validation Approach

The framework will be evaluated on four dimensions: **(1) Datasets**—MNIST and CIFAR-10 for vision tasks, FEMNIST for non-IID heterogeneity, MIMIC-III for healthcare applications; **(2) Attack scenarios**—Model inversion, membership inference, data poisoning, model poisoning, SMPC collusion; **(3) Baselines**—FedAvg, DP-FedAvg, Secure Aggregation, Krum, RFA; **(4) Metrics**—Accuracy under benign and adversarial conditions, privacy leakage rates, convergence speed, computational overhead. An expectation of 10–30% accuracy degradation with 90%+ attack mitigation; communication overhead of 50–100% is acceptable in domains where privacy and security are critical (healthcare, finance).

6. Conclusion

Current federated learning methods address only part of the threat landscape. This framework integrates Robust Aggregation, Secure Multi-Party Computation, and Differential Privacy within an adaptive control architecture. The result is resilience against sophisticated adversaries and support for legally compliant collaboration in privacy-sensitive domains such as GDPR-regulated finance and HIPAA-compliant healthcare.

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas.

Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[2] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[3] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[4] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020.

[5] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pages 634–643. PMLR, 2019.

[6] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[7] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[8] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[9] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pages 119–129, 2017.

[10] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

[11] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

[12] Jinhyun So, Başak Güler, and A Salman Avestimehr. Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications*, 39(7):2168–2181, 2021.