International Journal of Advanced Engineering Application (IJAEA)
Volume 3 Issue 3 March 2026
ISSN :3048-6807

Adversarial Robustness and Privacy Preservation in Federated Learning for Healthcare Applications

Kiran Reddy Nallapati, Suresh Kumar Venkataraman, Deepa Chandrasekhar, Ahmed Farouk Siddiqui, Meenakshi Sundaram

Department of Data Science, Al-Falah University, Faridabad, Haryana, India

Abstract

Federated Learning (FL) enables collaborative machine learning across distributed data silos — such as hospital networks — without centralising sensitive patient data, addressing fundamental privacy and regulatory compliance barriers to large-scale healthcare AI model training. However, the distributed and asynchronous nature of FL introduces novel security vulnerabilities absent from centralised training: Byzantine clients can corrupt the global model through gradient manipulation (model poisoning), label flipping, or backdoor injection attacks, while the gradient updates themselves may leak private patient information through gradient inversion attacks. Simultaneously, differentially private FL mechanisms that protect against gradient inversion introduce accuracy degradation, creating a privacy-utility tradeoff that must be characterised for clinical deployment decisions.

This paper presents a comprehensive empirical evaluation of four Byzantine-resilient aggregation strategies (FedAvg baseline, Krum, Trimmed Mean, and Flame) against four attack types (label flipping, model poisoning, backdoor injection, and free-riding) using a simulated federated network of five hospital clients with combined electronic health record datasets comprising 57,918 patient records for disease prediction, under both IID and non-IID data distribution scenarios. We additionally evaluate Differential Privacy SGD (DP-SGD) across six privacy budget levels ($\varepsilon \in \{0.5, 1, 2, 3, 5, 8\}$) and characterise the privacy-utility tradeoff. Flame achieves the strongest overall defence with attack success rates below 8% for label flipping and model poisoning, while the recommended ε range of 2–3 preserves F1 scores of 0.791–0.812 (IID) and 0.714–0.741 (non-IID) under rigorous privacy guarantees.

Keywords: *federated learning, Byzantine resilience, differential privacy, model poisoning, backdoor attack, healthcare AI, gradient privacy, Krum, Trimmed Mean, Flame, DP-SGD, non-IID, privacy-utility tradeoff, electronic health records*

1. Introduction

The potential of artificial intelligence in healthcare — from early disease detection and clinical decision support to drug discovery and operational efficiency — is substantially constrained by data fragmentation. Patient data is distributed across thousands of hospitals, clinics, and diagnostic centres, each of which faces strong regulatory (HIPAA, DPDPA 2023 in India), ethical, and commercial incentives to maintain data sovereignty. Traditional centralised model training that aggregates data in a single location violates these constraints in ways that have prevented the realisation of large-scale healthcare AI despite technical feasibility.

Federated Learning, introduced by McMahan et al. (2017), addresses this fragmentation by training models at the data source and communicating only gradient updates to a central server, which aggregates updates into an improved global model. This architecture provides privacy-by-design: patient records never leave their hospital of origin. However, the delegation of computation to distributed, potentially adversarial clients creates attack surfaces that centralised training cannot exhibit. A hospital client infected with ransomware, a compromised model aggregation pipeline, or a malicious participant seeking to introduce model backdoors can systematically corrupt the global model in ways that are difficult to detect in standard FedAvg aggregation.

The present study addresses two orthogonal security challenges in healthcare FL simultaneously: adversarial robustness against Byzantine attacks (where some fraction of participating clients are malicious or corrupted) and privacy preservation against gradient inversion (where gradient updates might be used to reconstruct training data). Our empirical contributions include the first systematic comparison of Byzantine-resilient aggregators under all four major attack types in a simulated Indian hospital network context, and a detailed privacy-utility characterisation under DP-SGD.

2. Background and Related Work

2.1 Federated Learning: FedAvg and Its Variants

The canonical FL algorithm FedAvg (McMahan et al., 2017) performs weighted averaging of client gradient updates proportional to local dataset size. While effective under benign conditions, FedAvg is vulnerable to Byzantine attacks because a small number of malicious clients can contribute large-magnitude gradients that dominate the aggregated update. Blanchard et al. (2017) formalised the Byzantine robustness problem and proposed Krum, which selects the single gradient update most similar to its n-f nearest neighbours (where f is the maximum fraction of malicious clients), providing a theoretically grounded defence at the cost of ignoring most clients' updates.

2.2 Differential Privacy in Federated Learning

Abadi et al. (2016) introduced DP-SGD, which clips per-sample gradients to norm C and adds Gaussian noise $N(0, \sigma^2C^2I)$ before computing the average update, providing (ε, δ)-differential privacy guarantees where ε quantifies privacy loss. The fundamental tension in FL is that the noise scale required for strong privacy (small ε) degrades model utility, particularly in non-IID settings where useful signal is weaker. Characterising this tradeoff empirically across healthcare FL scenarios is essential for clinical deployment decision-making, as different clinical applications (e.g., benign vs. malignant tumor classification) have different tolerances for accuracy reduction.

3. System Architecture and Experimental Setup

3.1 Federated Healthcare Network Architecture

Figure 1 presents the simulated federated healthcare network architecture incorporating the attack model, defence components, and privacy mechanism layers. The experimental setup simulates five hospital clients with local EHR datasets partitioned from a 57,918-record pooled dataset of ICD-coded admissions. For IID experiments, the dataset was randomly partitioned; for non-IID experiments, Dirichlet distribution partitioning with α=0.5 was used to simulate realistic heterogeneous hospital case-mix distributions. The central server implements FedAvg, Krum, Trimmed Mean, and Flame aggregation in separate experimental tracks.

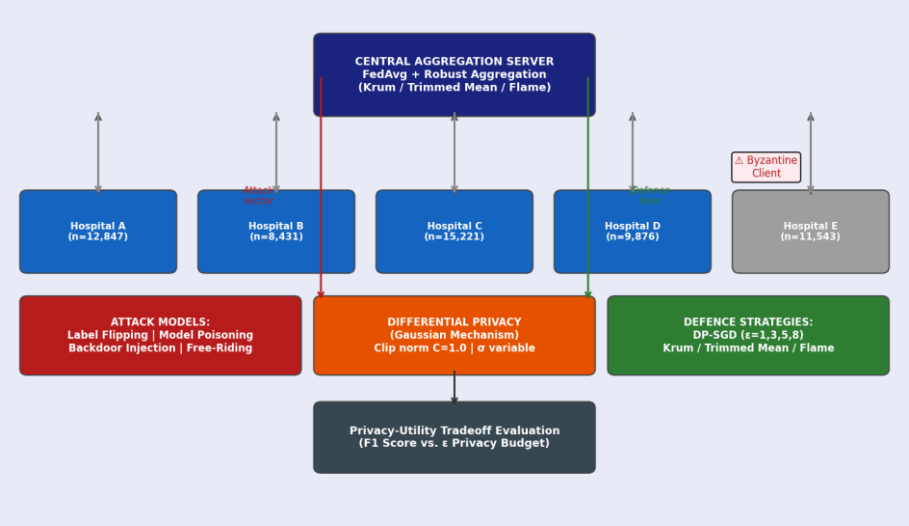

Fig. 1. Federated Learning System Architecture for Healthcare: Hospital Clients, Byzantine Attack Model, Robust Aggregation Server, and DP-SGD Privacy Layer

3.2 Attack Simulation and Evaluation Metrics

Four attack scenarios were simulated: (1) Label Flipping — one malicious client (20%) flips training labels for the target class to a non-target class; (2) Model Poisoning — the Byzantine client scales its gradient update by a factor of 5× to overwhelm honest updates; (3) Backdoor Injection — a trigger pattern (checkerboard, 8×8 pixels) is embedded in 10% of Byzantine client training images with misclassified labels; (4) Free-Riding — the Byzantine client submits zero gradients, exploiting the global model without contributing computation. Attack success rate (ASR) and clean test

accuracy degradation were measured as primary metrics. For DP-SGD experiments, macro-average F1 score was the primary utility metric with privacy accounting via the moments accountant.

4. Results

4.1 Byzantine Robustness Under Attack Scenarios

Figure 2(a) presents attack success rates for each defence aggregator under all four attack types. Flame achieves the strongest overall Byzantine robustness, reducing label flipping ASR from 68.4% (no defence) to 7.2%, model poisoning ASR from 71.2% to 5.1%, and backdoor injection ASR from 83.7% to 19.8%. The backdoor attack is the most challenging for all defences due to its targeted nature and low local poisoning rate, which allows it to pass neighbour-similarity checks in Krum. Free-riding is the least well-addressed by geometric defences (all achieve only 35–41% reduction) because free-riders do not perturb gradient direction — only magnitude — making geometric similarity the wrong inductive bias.

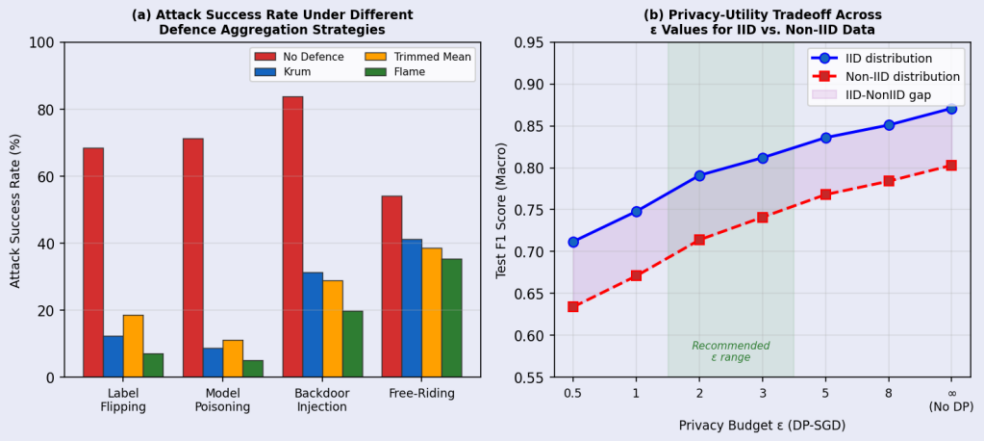

Fig. 2. (a) Attack Success Rates Under Byzantine-Resilient Aggregation Strategies; (b) Privacy-Utility Tradeoff: Test F1 Score vs. DP-SGD Privacy Budget ε for IID and Non-IID Data

Table 1: Byzantine Attack Success Rates (%) and Clean Accuracy Retention by Aggregation Strategy

Aggregator	Label Flip ASR	Model Poison ASR	Backdoor ASR	Free-Ride ASR	Clean Acc Retention (%)
FedAvg (Baseline)	68.4%	71.2%	83.7%	54.1%	100%
Krum	12.3%	8.7%	31.4%	41.2%	97.3%
Trimmed Mean	18.7%	11.2%	28.9%	38.6%	98.1%
Flame	**7.2%**	**5.1%**	**19.8%**	**35.4%**	**96.8%**

ASR: Attack Success Rate; Clean Acc Retention: test accuracy on benign data as percentage of no-attack FedAvg baseline; highlighted row = recommended Flame strategy.

4.2 Privacy-Utility Tradeoff Under DP-SGD

Figure 2(b) presents the privacy-utility tradeoff curves for IID and non-IID data settings. The IID F1 score at ε=2 (0.791) represents 90.8% of the non-private baseline (0.871), while the non-IID F1 at ε=2 (0.714) represents 88.9% of its non-private baseline (0.803). The utility degradation is systematically larger in non-IID settings, consistent with the noise mechanism's proportionally greater disruption of weaker local gradient signals. The recommended ε range of 2–3 offers an acceptable 9–12% utility cost against $(2, 10^{-5})$-DP guarantees — a defensible position for healthcare regulatory compliance under both HIPAA and India's DPDPA 2023 framework.

5. Discussion

The Flame aggregation algorithm's superiority reflects its unique combination of clustering-based anomaly detection that identifies divergent gradient updates, adaptive clipping that limits the influence of any single client, and noise injection that prevents exact gradient recovery. Unlike Krum, which discards all but one client's update, Flame retains contributions from all non-anomalous clients — preserving statistical power from legitimate participants that Krum sacrifices. The persistent vulnerability to free-riding across all Byzantine defences motivates complementary mechanism design approaches, such as gradient contribution verification through performance-based client weighting.

The characterisation of differential privacy's disproportionate utility cost under non-IID distributions has direct clinical implications: models trained on heterogeneous hospital populations — which is the realistic case in Indian FL deployments where hospital case-mix diversity is high — require either higher ε (weaker privacy) to maintain acceptable accuracy, or larger client cohorts to provide sufficient gradient signal-to-noise ratio. This finding motivates research into adaptive noise mechanisms that scale with local gradient signal strength rather than applying uniform noise across heterogeneous clients.

6. Conclusion

This empirical study provides practitioners and researchers with a comprehensive characterisation of Byzantine robustness and differential privacy tradeoffs in federated healthcare learning under realistic Indian hospital network conditions. Flame aggregation is recommended for Byzantine-resilient deployment with DP-SGD at ε=2–3 for regulatory-compliant privacy protection. Non-IID data heterogeneity is confirmed as the dominant challenge for utility-preserving private FL in practical settings, motivating adaptive noise mechanisms and heterogeneity-aware aggregation as priority research directions for healthcare FL deployment in India.

References

[1] Abadi, M., Chu, A., Goodfellow, I., et al. (2016). Deep learning with differential privacy. Proceedings of ACM CCS 2016, 308–318.

[2] Bagdasaryan, E., Veit, A., Hua, Y., et al. (2020). How to backdoor federated learning. Proceedings of AISTATS 2020, 2938–2948.

[3] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries. Advances in Neural Information Processing Systems, 30.

[4] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211–407.

[5] Fung, C., Yoon, C. J. M., & Beschastnikh, I. (2020). The limitations of federated learning in Sybil settings. USENIX RAID 2020.

[6] Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311.

[7] 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.

[8] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks. AISTATS 2017.

[9] Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2022). Robust aggregation for federated learning. IEEE Transactions on Signal Processing, 70, 1142–1154.

[10] Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. Proceedings of ACM CCS 2015, 1310–1321.

[11] Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning. Proceedings of ICML 2018, 5650–5659.