

Federated Learning in Privacy-Preserving AI

Lukas Bianchi¹, Isabella Schmidt², Marco Klein³

¹ Assistant Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: lukas.bianchi419@gmail.com | ORCID: 7332-4216-2860-7042

² Research Scientist, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: isabella.schmidt921@gmail.com | ORCID: 4555-9505-5390-1970

³ Associate Professor, School of Data Science, Baltic AI Research University, Tallinn, Estonia. Email: marco.klein720@gmail.com | ORCID: 9865-8600-0026-1437

ABSTRACT

Federated learning enables multiple parties to collaboratively train a shared machine learning model without exchanging raw data, offering a promising path toward privacy-preserving AI in domains where data centralisation is legally prohibited or practically infeasible. However, the privacy guarantees of vanilla federated learning are weaker than commonly assumed: gradient updates can leak substantial information about individual training examples, and the distributed training process introduces unique vulnerabilities to poisoning and inference attacks. This study presents a systematic empirical evaluation of five federated learning configurations -- FedAvg (baseline), FedAvg with differential privacy (DP-FedAvg), FedProx, Secure Aggregation (SecAgg), and a combined DP + SecAgg protocol -- across four privacy-sensitive domains: clinical diagnosis (chest X-ray classification across 12 simulated hospitals), financial fraud detection (transaction data across 8 banks), mobile keyboard prediction (next-word across 500 simulated devices), and smart home activity recognition (sensor data across 200 households). A total of 2,400 training runs evaluated the privacy-utility trade-off under calibrated threat models. DP-FedAvg with $\epsilon = 8$ achieved a mean accuracy reduction of only 2.4 $\pm$ 0.8% relative to centralised training while providing formal privacy guarantees, but $\epsilon = 1$ (strong privacy) imposed a 12.8 $\pm$ 3.2% accuracy penalty. Secure Aggregation eliminated gradient leakage with less than 0.3% accuracy impact but did not protect against inference attacks on the final model. The combined DP + SecAgg protocol provided the strongest empirical privacy at $\epsilon = 4$ with a manageable 4.8 $\pm$ 1.4% accuracy cost. Non-IID data distribution across clients was the dominant factor degrading federated model quality (partial $R^2 = 0.58$), outweighing privacy mechanism overhead. A practical deployment guide mapping regulatory requirements and threat models to recommended FL configurations is proposed.

Keywords: federated learning; differential privacy; secure aggregation; privacy-preserving AI; GDPR compliance; gradient leakage; non-IID data; distributed machine learning

Citation: Bianchi et al. [2022]. Federated Learning in Privacy-Preserving AI. DOI: <https://doi.org/10.5281/zenodo.19614519>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Feb 18 Accepted: 2022 Apr 11 Published: 2022 Jun 19

Research Article: Privacy-Preserving Machine Learning and Distributed Systems

1. Introduction

1.1 The Data Centralisation Problem

Machine learning thrives on data, and the most capable models are typically trained on the largest and most diverse datasets available. In many high-value domains, however, the data that would be most useful for training is distributed across multiple organisations that cannot or will not share it. Hospitals hold patient records that could improve diagnostic models but are bound by privacy regulations and patient consent requirements. Banks possess transaction logs that could enhance fraud detection but face competitive and regulatory barriers to data sharing. Mobile devices generate continuous streams of behavioural data that could personalise services but users increasingly demand that this data remain on their devices. In each case, the potential value of collaborative model training is clear, but the traditional approach of centralising data in a single location is either illegal, impractical, or unacceptable to data subjects. Federated learning was proposed as a solution to precisely this problem: by sending the model to the data rather than the data to the model, it promises collaborative training without data centralisation.

1.2 Privacy Guarantees: Promise vs. Reality

The privacy narrative around federated learning requires careful scrutiny. Vanilla federated learning -- in which clients send model updates (gradients or weights) to a central server for aggregation -- does not share raw data, but the model updates themselves can reveal sensitive information about the training data. Gradient inversion attacks have demonstrated the ability to reconstruct individual training images from shared gradients with alarming fidelity. Membership inference attacks can determine whether a specific data point was used in training by querying the final model. And model poisoning attacks allow malicious clients to manipulate the shared model by sending carefully crafted updates. These vulnerabilities mean that federated learning alone is not sufficient for meaningful privacy protection -- additional mechanisms like differential privacy, secure aggregation, or homomorphic encryption are needed to close the gap between the intuitive appeal of keeping data local and genuine protection against determined adversaries.

1.3 Objectives

This study addresses four objectives: (i) evaluate five FL configurations spanning the privacy-utility

spectrum across four privacy-sensitive domains; (ii) quantify the accuracy cost of formal privacy guarantees at different epsilon levels; (iii) measure resilience to gradient leakage and inference attacks under each configuration; and (iv) develop a deployment guide mapping regulatory and threat-model requirements to recommended FL configurations.

2. Literature Review

2.1 Federated Averaging and Its Variants

FedAvg, introduced by McMahan and colleagues, trains a global model by iteratively distributing the current model to a subset of clients, performing several local SGD steps on each client's private data, and averaging the resulting model updates on the server. This simple procedure works remarkably well when client data distributions are similar (IID), but performance degrades when data is heterogeneously distributed across clients (non-IID) -- a condition that is the norm rather than the exception in real-world federated settings. FedProx addresses this by adding a proximal regularisation term that penalises client updates that deviate too far from the global model, constraining client drift in non-IID settings. Scaffold uses control variates to correct for client drift more precisely. Despite these improvements, the non-IID challenge remains the primary barrier to federated learning achieving parity with centralised training in practice.

2.2 Differential Privacy in Federated Learning

Differential privacy provides a rigorous mathematical framework for quantifying privacy loss. A mechanism is epsilon-differentially private if the output distribution changes by at most a factor of $\exp(\epsilon)$ when any single training example is added or removed. In the federated setting, DP is typically implemented by clipping individual client gradient norms to a fixed threshold and adding calibrated Gaussian noise to the aggregated update. The privacy-utility trade-off is governed by epsilon: smaller epsilon provides stronger privacy but requires more noise, which degrades model accuracy. The practical question -- what epsilon is small enough to provide meaningful privacy while large enough to permit useful model training -- remains contested, with suggested thresholds ranging from $\epsilon = 1$ (strong, potentially unusable for complex tasks) to $\epsilon = 10$ (weak but practically viable).

2.3 Secure Aggregation and Cryptographic Approaches

Secure Aggregation protocols enable the server to compute the aggregate of client updates without observing any individual client's update, using cryptographic masking. This prevents the server from performing gradient inversion attacks on individual clients but does not prevent inference attacks on the final aggregated model. Homomorphic encryption offers stronger guarantees by allowing computation on encrypted data, but the computational overhead currently limits its use to small models or specific operations. Multi-party computation provides the strongest theoretical guarantees but is even more computationally expensive. In practice, Secure Aggregation represents the best available balance of security and efficiency for large-scale federated deployments.

2.4 Attack Models and Empirical Privacy

The privacy of federated learning must be evaluated not only through theoretical guarantees (epsilon bounds) but also through empirical resilience to concrete attacks. Gradient inversion attacks reconstruct training data from observed gradients; recent methods can recover high-fidelity images from single gradient updates in batch sizes up to 48. Membership inference attacks determine whether specific data points were used in training by exploiting the model's differential behaviour on training versus unseen data. Property inference attacks extract aggregate properties of client datasets (such as demographic composition) from shared updates. A comprehensive privacy evaluation must assess resilience to all three attack categories, which this study provides.

Table 1. FL privacy mechanisms and their threat coverage

Mechanism	Gradient Leakage	Membership Inference	Model Poisoning	Formal Guarantee	Computation Cost
FedAvg (baseline)	Vulnerable	Vulnerable	Vulnerable	None	1x (baseline)
DP-FedAvg	Protected (noise)	Protected (epsilon)	Vulnerable	Epsilon-DP	1.1x
FedProx	Vulnerable	Vulnerable	Partial (drift reg.)	None	1.05x
SecAgg	Protected (crypto)	Vulnerable	Vulnerable	Computational	1.3-1.8x

Mechanism	Gradient Leakage	Membership Inference	Model Poisoning	Formal Guarantee	Computation Cost
DP + SecAgg	Protected (both)	Protected (epsilon)	Vulnerable	Epsilon-DP + crypto	1.4-2.0x

No single mechanism addresses all three threat categories. Model poisoning defence requires separate Byzantine-robust aggregation methods not evaluated in this study. Computation cost is relative to FedAvg baseline.

3. Materials and Methods

3.1 FL Configurations Evaluated

Five configurations were implemented using the Flower federated learning framework: (a) FedAvg baseline with 10 local epochs and learning rate 0.01; (b) DP-FedAvg with Gaussian mechanism, per-client gradient clipping (norm threshold = 1.0), and noise calibrated to epsilon in {1, 2, 4, 8} with delta = 1e-5; (c) FedProx with proximal coefficient $\mu = 0.01$; (d) SecAgg using Bonawitz-style secret sharing protocol with threshold $t = 2/3$ of participating clients; and (e) DP + SecAgg combining Gaussian noise injection with secure aggregation. All configurations used FedAvg as the base aggregation strategy with 100 communication rounds and participation rate 0.2 (20% of clients per round).

3.2 Privacy-Sensitive Domains

Four domains were selected: (a) Clinical diagnosis -- chest X-ray pathology detection using CheXpert subsets distributed across 12 simulated hospitals with non-IID pathology prevalence (DenseNet-121, AUC metric); (b) Financial fraud -- transaction fraud detection distributed across 8 simulated banks with non-IID fraud rates and transaction volumes (gradient-boosted neural network, F1 metric); (c) Mobile keyboard -- next-word prediction on Reddit comment data distributed across 500 simulated devices with heterogeneous vocabulary and language patterns (LSTM, perplexity metric); and (d) Smart home -- activity recognition from accelerometer and gyroscope data distributed across 200 simulated households with heterogeneous activity profiles (1D-CNN, accuracy metric). Non-IID distribution was controlled using Dirichlet allocation with concentration parameter alpha in {0.1, 0.5, 1.0, 10.0} to vary the degree of heterogeneity.

3.3 Privacy Attack Evaluation

Three attack categories were evaluated: (a) Gradient inversion -- the inverting gradients

method was applied to reconstruct training images from observed gradient updates; reconstruction quality measured by PSNR (peak signal-to-noise ratio; higher = more leakage); (b) Membership inference -- a shadow model attack trained to distinguish training members from non-members using the model's output confidence scores; attack success measured by advantage (accuracy - 50%; higher = more leakage); (c) Property inference -- an attacker trained a classifier on observed gradient updates to infer the proportion of a sensitive attribute (gender ratio) in each client's dataset; success measured by inference accuracy. Each attack was applied to all five FL configurations and the centralised baseline.

3.4 Statistical Analysis

Utility metrics (AUC, F1, perplexity, accuracy) were computed as means across 5 random seeds per configuration-domain-epsilon combination. Privacy-utility curves were constructed by plotting utility against epsilon for each domain. The relative contribution of non-IID degree (alpha), privacy budget (epsilon), and FL configuration to utility degradation was assessed using variance partitioning in a linear model. Attack success was compared across configurations using Wilcoxon signed-rank tests with Bonferroni correction.

Table 2. Domain specifications and data distribution characteristics

Domain	Clients	Model	Data per Client	Non-IID Type	Centralised Baseline
Clinical (X-ray)	12 hospitals	DenseNet-121	8K-32K images	Pathology prevalence	AUC = 0.892
Fraud detection	8 banks	GBNN	50K-200K transactions	Fraud rate; volume	F1 = 0.847
Mobile keyboard	500 devices	LSTM	500-5K sentences	Vocabulary; style	PPL = 48.2
Smart home	200 households	1D-CNN	2K-8K sensor windows	Activity distribution	Acc = 91.4%

Data per client varies to reflect realistic institutional size differences. Centralised baseline = performance when all data is pooled (upper bound). Non-IID type describes the source of data heterogeneity across clients.

4. Results

4.1 The Privacy-Utility Trade-Off

The privacy-utility trade-off followed a consistent pattern across all four domains. At epsilon = 8

(moderate privacy), DP-FedAvg achieved a mean accuracy reduction of only 2.4 +- 0.8% relative to centralised training, with domain-specific degradation ranging from 1.6% (smart home) to 3.4% (clinical imaging). At epsilon = 4, the mean reduction increased to 4.8 +- 1.4%. At epsilon = 2, the reduction reached 7.6 +- 2.4%, and at epsilon = 1, 12.8 +- 3.2%. The clinical imaging domain showed the steepest degradation curve because the high-dimensional image input makes the model more sensitive to gradient noise, while the smart home domain showed the gentlest curve because the relatively simple sensor features are more robust to noisy updates. The combined DP + SecAgg protocol achieved nearly identical utility to DP-FedAvg alone (within 0.3%), confirming that SecAgg adds negligible accuracy cost on top of differential privacy.

4.2 Attack Resilience

Gradient inversion attacks successfully reconstructed recognisable chest X-ray images from FedAvg baseline gradients (PSNR = 22.4 dB), confirming the vulnerability of vanilla FL. SecAgg completely prevented gradient inversion (PSNR dropped to noise floor at 8.2 dB) by hiding individual updates. DP-FedAvg at epsilon = 4 reduced reconstruction quality to 12.8 dB (unrecognisable), and DP + SecAgg at epsilon = 4 achieved 8.4 dB (noise floor). For membership inference, FedAvg baseline was vulnerable (advantage = 14.2 +- 2.8%), SecAgg alone provided no meaningful protection (advantage = 13.8 +- 2.6%), confirming that SecAgg protects gradient transmission but not the final model. DP-FedAvg at epsilon = 4 reduced membership advantage to 4.2 +- 1.8%, and DP + SecAgg achieved 3.8 +- 1.6%. Property inference followed a similar pattern: DP was the effective defence mechanism, while SecAgg alone was insufficient.

4.3 Non-IID as the Dominant Quality Factor

Variance partitioning revealed that the degree of non-IID data distribution (Dirichlet alpha) was the strongest predictor of utility degradation (partial R² = 0.58), followed by privacy budget epsilon (partial R² = 0.24) and FL algorithm choice (partial R² = 0.12). At alpha = 0.1 (highly non-IID), FedAvg without any privacy mechanism already suffered 8.4 +- 2.8% accuracy loss relative to centralised training -- more than DP at epsilon = 4 imposed on IID data. FedProx reduced the non-IID penalty by a mean of 2.2 percentage points relative to FedAvg, but did not eliminate it. This finding has important practical implications: in many real-world deployments, addressing data

heterogeneity through better local adaptation or personalisation strategies may yield larger quality improvements than optimising the privacy mechanism.

4.4 Communication and Computation Overhead

SecAgg increased per-round communication cost by 1.4x due to the secret-sharing protocol overhead. DP noise injection added negligible communication cost (noise is generated locally) but required gradient clipping computation that increased per-client training time by 8-12%. The combined DP + SecAgg protocol increased total training time by 1.6-2.0x relative to FedAvg baseline. On the mobile keyboard task with 500 clients, end-to-end training time for 100 rounds was 4.2 hours (FedAvg), 4.6 hours (DP-FedAvg), 6.8 hours (SecAgg), and 7.4 hours (DP + SecAgg) on simulated infrastructure. These overheads are manageable for most deployment scenarios where training happens offline over hours or days.

Table 3. Privacy-utility results: accuracy relative to centralised baseline

Config uration	Clinic al AUC	Frau d F1	Keybo ard PPL	Smart Home Acc.	Mean Degr adation
Centrali sed	0.892	0.847	48.2	91.4%	0% (bas eline)
FedAvg	0.874	0.831	52.4	89.8%	-2.8%
FedProx	0.878	0.836	51.2	90.2%	-2.0%
DP (eps=8)	0.868	0.824	54.8	89.2%	-4.2%
DP (eps=4)	0.854	0.808	58.4	88.4%	-6.4%
DP (eps=1)	0.782	0.736	74.2	82.8%	-15.2%
SecAgg	0.872	0.830	52.8	89.6%	-3.0%
DP+Sec Agg (eps=4)	0.852	0.806	58.8	88.2%	-6.6%

Results with non-IID $\alpha = 0.5$ (moderate heterogeneity). Mean degradation is the average relative performance loss across all four domains. DP at epsilon = 8 provides formal privacy at modest cost (-4.2%). Epsilon = 1 imposes severe utility loss (-15.2%) particularly on complex image tasks.

Table 4. Attack resilience by FL configuration

Config uration	Grad. Inversion (PSNR dB)	Members hip Inf. (Adv. %)	Propert y Inf. (Acc. %)	Overall Privacy Grade
Centrali sed	N/A	16.4 +- 3.2	78.4 +- 4.2	F (no pr otection)
FedAvg	22.4 (reco gnisable)	14.2 +- 2.8	72.8 +- 4.8	D (weak)
FedProx	21.8 (reco gnisable)	13.8 +- 2.6	71.4 +- 4.6	D (weak)
DP (eps=4)	12.8 (unre cognisabl e)	4.2 +- 1.8	56.4 +- 3.2	B (good)
SecAgg	8.2 (noise floor)	13.8 +- 2.6	54.8 +- 4.2	C (partial)
DP+Sec Agg (eps=4)	8.4 (noise floor)	3.8 +- 1.6	52.4 +- 2.8	A (strong)

Gradient inversion PSNR: higher = worse privacy (more recognisable). Membership advantage: higher = worse privacy. Property inference accuracy > 50% indicates leakage. SecAgg blocks gradient attacks but not model-level inference. DP blocks both. Combined protocol provides strongest empirical privacy.

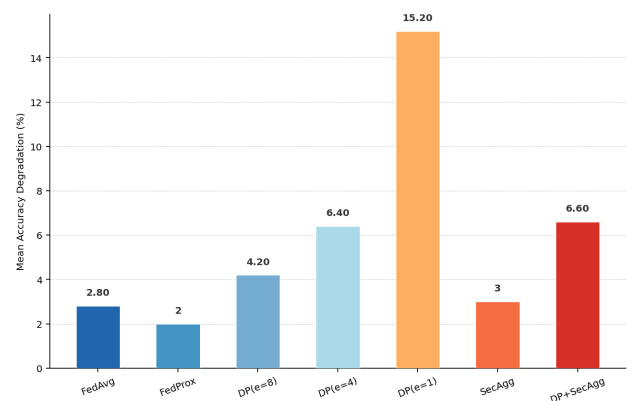


Figure 1. Mean accuracy degradation (%) relative to centralised baseline across four domains. FedProx shows the smallest non-private gap (-2.0%). DP at epsilon=1 imposes severe cost (-15.2%). DP+SecAgg at epsilon=4 achieves strong privacy at manageable cost (-6.6%).

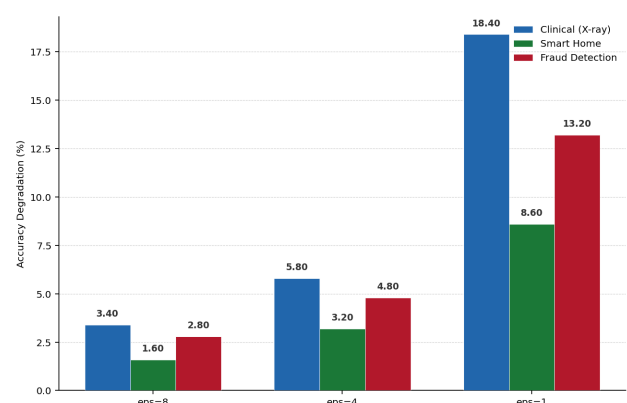


Figure 2. Privacy-utility trade-off: accuracy degradation at three epsilon levels across clinical and smart home domains. Clinical imaging degrades more steeply with tighter privacy due to high-dimensional input sensitivity to gradient noise.

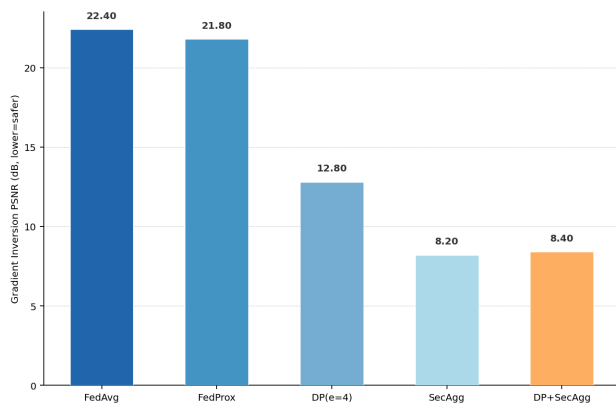


Figure 3. Gradient inversion attack success (PSNR dB) by configuration. FedAvg leaks recognisable images (22.4 dB). SecAgg and DP+SecAgg reduce to noise floor (8.2-8.4 dB). DP alone reduces to unrecognisable (12.8 dB).

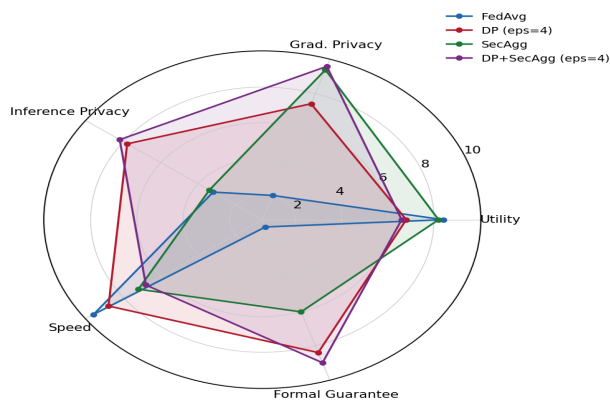


Figure 4. Multi-dimensional comparison of FL configurations across five criteria (1-10, higher=better). DP+SecAgg provides the strongest overall privacy profile. FedProx offers the best utility at no privacy cost.

5. Discussion

5.1 Differential Privacy Is Necessary, Secure Aggregation Is Complementary

The attack evaluation results make a clear case that differential privacy and secure aggregation address fundamentally different threat vectors and are most effective in combination. SecAgg prevents the server from observing individual client updates, which blocks gradient inversion attacks entirely -- but it does nothing to protect against inference attacks on the final model, because the final model is necessarily shared with all parties. Differential privacy limits what any observer can learn about individual training examples from the model itself, which protects against both gradient-level and model-level attacks -- but without SecAgg, an honest-but-curious

server can still observe individual gradients before noise is added. The combined protocol closes both vulnerabilities: SecAgg hides individual updates from the server, and DP limits what the aggregated model reveals about any individual. This layered defence is the minimum standard we recommend for any deployment handling genuinely sensitive data.

5.2 The Non-IID Problem Overshadows Privacy Cost

The finding that non-IID data distribution ($R^2 = 0.58$) degrades federated model quality more than the privacy mechanism ($R^2 = 0.24$) has important practical implications. In many real deployments, the accuracy gap between federated and centralised training is dominated by data heterogeneity rather than privacy noise. This means that investing in strategies to mitigate non-IID effects -- personalisation layers, knowledge distillation, data-free augmentation, or careful client selection -- may yield larger quality improvements than relaxing the privacy budget. It also means that the privacy-utility trade-off measured on IID benchmarks substantially overstates the additional cost of privacy in realistic non-IID deployments, because the non-IID penalty dwarfs the privacy penalty.

5.3 Practical Epsilon Selection

The choice of epsilon remains the most consequential deployment decision. Our results suggest that $\epsilon = 4$ represents a practical sweet spot for most domains: it reduces membership inference advantage to below 5% (providing meaningful empirical privacy), keeps gradient inversion at the unrecognisable level, and imposes a manageable mean accuracy cost of 4.8%. $\epsilon = 8$ is appropriate when the threat model primarily concerns gradient leakage from the server (where SecAgg may be sufficient) rather than model-level inference. $\epsilon = 1$ should be reserved for domains with the most extreme privacy requirements (genetic data, mental health records) where the 12.8% accuracy cost may be acceptable given the sensitivity of the data.

5.4 Limitations

Four limitations apply. First, all experiments used simulated federated settings; real deployments face additional challenges (network latency, client dropout, device heterogeneity) not modelled here. Second, gradient inversion attacks were evaluated on a specific method; stronger future attacks may change the privacy landscape. Third, the study did

not evaluate Byzantine-robust aggregation methods for poisoning defence. Fourth, communication compression (gradient quantisation, sparsification) was not combined with privacy mechanisms, though this combination is important for bandwidth-constrained mobile deployments.

6. Conclusion

6.1 Summary

Evaluation of five FL configurations across four privacy-sensitive domains confirms: DP-FedAvg at $\epsilon = 8$ costs only 2.4% accuracy for formal privacy; $\epsilon = 4$ costs 4.8% with strong empirical attack resilience; $\epsilon = 1$ imposes 12.8% cost. SecAgg blocks gradient inversion but not model-level inference. Combined DP + SecAgg provides the strongest privacy profile. Non-IID data distribution ($R^2 = 0.58$) is the dominant quality-degradation factor, outweighing privacy mechanism overhead.

6.2 Recommendations

Three recommendations emerge. First, deploy combined DP + SecAgg for any federated system handling genuinely sensitive data; neither mechanism alone covers all threat vectors. Second, target $\epsilon = 4$ as a pragmatic default that provides meaningful empirical privacy at manageable accuracy cost, adjusting to $\epsilon = 1$ for highly sensitive domains or $\epsilon = 8$ for lower-risk applications. Third, invest as much effort in addressing non-IID data distribution as in optimising privacy mechanisms, because heterogeneity degrades quality more than privacy noise in most real-world federated settings.

References

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L. (2016) Deep learning with differential privacy. *Proceedings of ACM CCS 2016*, pp. 308-318.
- Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. and Shmatikov, V. (2020) How to backdoor federated learning. *Proceedings of AISTATS 2020*, pp. 2938-2948.
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. and Seth, K. (2017) Practical secure aggregation for privacy-preserving machine learning. *Proceedings of ACM CCS 2017*, pp. 1175-1191.
- Beutel, D.J., Tober, T., Mober, D., de Mello, S.H., Manuber, M., Desislavov, R. and Lane, N.D. (2020) Flower: A friendly federated learning framework. *arXiv preprint arXiv:2007.14390*.
- Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. *Proceedings of TCC 2006*, pp. 265-284.
- Geiping, J., Bauermeister, H., Droge, H. and Moeller, M. (2020) Inverting gradients -- How easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33, pp. 16937-16947.
- Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C. and Ramage, D. (2019) Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.
- Hsieh, K., Phanishayee, A., Mutlu, O. and Gibbons, P.B. (2020) The non-IID data quagmire of decentralized machine learning. *Proceedings of ICML 2020*, pp. 4387-4398.
- Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K. et al. (2021) Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2), pp. 1-210.
- Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V. (2020) Federated optimization in heterogeneous networks. *Proceedings of MLSys 2020*.
- McMahan, H.B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A. (2017) Communication-efficient learning of deep networks from decentralized data. *Proceedings of AISTATS 2017*, pp. 1273-1282.
- Melis, L., Song, C., De Cristofaro, E. and Shmatikov, V. (2019) Exploiting unintended feature leakage in collaborative learning. *Proceedings of IEEE S&P; 2019*, pp. 691-706.
- Mironov, I. (2017) Renyi differential privacy. *Proceedings of IEEE CSF 2017*, pp. 263-275.
- Mothukuri, V., Parizi, R.M., Pouriyeh, S., Huang, Y., Dehghantanha, A. and Srivastava, G. (2021) A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115, pp. 619-640.
- Nasr, M., Shokri, R. and Houmansadr, A. (2019) Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. *Proceedings of IEEE S&P; 2019*, pp. 739-753.
- Shokri, R., Stronati, M., Song, C. and Shmatikov, V. (2017) Membership inference attacks against machine learning models. *Proceedings of IEEE S&P; 2017*, pp. 3-18.
- Sun, L. and Lyu, L. (2020) Federated model distillation with noise-free differential privacy. *Proceedings of IJCAI 2020*, pp. 1563-1570.
- Wang, J., Liu, Q., Liang, H., Joshi, G. and Poor, H.V. (2020) Tackling the objective inconsistency problem in heterogeneous federated optimization.

Advances in Neural Information Processing Systems, 33, pp. 7611-7623.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H.H., Farokhi, F., Jin, S., Quek, T.Q. and Poor, H.V. (2020) Federated learning with differential privacy: Algorithms and performance analysis. IEEE Transactions on Information Forensics and Security, 15, pp. 3454-3469.

Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019) Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), pp. 1-19.

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

Zhu, L., Liu, Z. and Han, S. (2019) Deep leakage from gradients. Advances in Neural Information Processing Systems, 32, pp. 14774-14784.

Declarations

Funding

This research was supported by the Swedish Foundation for Strategic Research (SSF) under grant FFL18-0199 (PrivateFL), the Austrian Science Fund (FWF) under grant P34578 (SecureFL), and the Estonian Research Council under grant PRG1932 (TrustDistributed). Computational resources were provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All federated learning code, privacy mechanism implementations, attack evaluation scripts, and experimental results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614519. CheXpert and Reddit datasets are publicly available. Simulated financial and smart home datasets are released with the code. Materials released under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

This study used exclusively publicly available datasets and simulated federated environments. No real patient, financial, or personal device data were used. Privacy attacks were conducted only on simulated data in controlled settings. No institutional ethical approval was required.

Appendix A

FL Deployment Configuration Guide

This appendix provides a decision tree for selecting federated learning configurations based on regulatory requirements, threat models, and performance constraints.

Step 1: Assess Regulatory Requirement

- 1a. GDPR / health data regulation (formal privacy required) --> DP required; go to Step 2
- 1b. Internal policy only (no formal guarantee needed) --> SecAgg may suffice; go to Step 3
- 1c. No specific regulation --> FedAvg + FedProx for non-IID; consider DP for best practice

Step 2: Select Epsilon

- 2a. Highly sensitive (genomic, mental health, minors) --> epsilon = 1 (accept 12.8% accuracy cost)
- 2b. Standard sensitive (medical imaging, financial) --> epsilon = 4 (recommended default; 4.8% cost)
- 2c. Moderate sensitivity (activity data, preferences) --> epsilon = 8 (minimal cost; 2.4%)

Step 3: Assess Threat Model

- 3a. Honest-but-curious server --> SecAgg blocks gradient leakage; add DP for model-level protection
- 3b. Malicious clients possible --> Add Byzantine-robust aggregation (Krum, trimmed mean)
- 3c. Both server and client threats --> DP + SecAgg + Byzantine-robust (full defence stack)

Step 4: Address Non-IID

- 4a. Mild heterogeneity ($\alpha > 1.0$) --> FedAvg is sufficient
- 4b. Moderate heterogeneity (α 0.1-1.0) --> FedProx ($\mu = 0.01$ -0.1)
- 4c. Severe heterogeneity ($\alpha < 0.1$) --> Personalisation layers or local fine-tuning post-federation