Empirical Analysis of Privacy-Fairness-Accuracy Trade-offs in Federated Learning: A Step Towards Responsible AI

Dawood Wasif[1], Dian Chen[1], Sindhuja Madabushi[1], Nithin Alluru[1], Terrence J. Moore[2], Jin-Hee Cho[1]

[1]Virginia Tech, Blacksburg, Virginia, U.S.A.
[2]US Army Research Laboratory, Adelphi, Maryland, U.S.A.
{dawoodwasif, dianc, msindhuja, nithin, jicho}@vt.edu, terrence.j.moore.civ@army.mil

Abstract

Federated Learning (FL) enables collaborative model training while preserving data privacy; however, balancing privacy preservation (PP) and fairness poses significant challenges. In this paper, we present the first unified large-scale empirical study of privacy–fairness–utility trade-offs in FL, advancing toward responsible AI deployment. Specifically, we systematically compare Differential Privacy (DP), Homomorphic Encryption (HE), and Secure Multi-Party Computation (SMC) with fairness-aware optimizers including q-FedAvg, q-MAML, Ditto, evaluating their performance under IID and non-IID scenarios using benchmark (MNIST, Fashion-MNIST) and real-world datasets (Alzheimer's MRI, credit-card fraud detection). Our analysis reveals HE and SMC significantly outperform DP in achieving equitable outcomes under data skew, although at higher computational costs. Remarkably, we uncover unexpected interactions: DP mechanisms can negatively impact fairness, and fairness-aware optimizers can inadvertently reduce privacy effectiveness. We conclude with practical guidelines for designing robust FL systems that deliver equitable, privacy-preserving, and accurate outcomes.

Code — https://github.com/dawoodwasif/Privacy_vs_Fairness_vs_Accuracy

Extended version — https://arxiv.org/abs/2503.16233

Introduction

Federated Learning (FL) enables decentralized model training while preserving privacy, yet its deployment in realistic scenarios such as healthcare and finance raises challenges at the intersection of privacy, fairness, and real-world impact. Privacy-preserving techniques like Differential Privacy (DP), Homomorphic Encryption (HE), and Secure Multi-Party Computation (SMC) enhance security but introduce fairness concerns. DP's noise protects data but degrades utility for underrepresented groups, exacerbating bias (Bagdasaryan, Poursaeed, and Shmatikov 2019). HE and SMC preserve data integrity and mitigate fairness disparities but impose additional computational costs, potentially limiting participation (Truex et al. 2019; Xu et al. 2023). Despite advancements, privacy-fairness trade-offs in FL remain underexplored, necessitating privacy-aware fairness mechanisms that align with responsible AI to ensure equitable and transparent model performance.

A critical but often overlooked issue is *client fairness*, ensuring equitable model performance across clients with heterogeneous data and resources. Unlike algorithmic fairness, which addresses bias in decisions, client fairness examines FL's impact on diverse clients. In dermatology, FL models trained on imbalanced client data favored overrepresented skin types, reducing accuracy for underrepresented groups (Weng et al. 2020), highlighting the need for bias mitigation in healthcare AI.

To promote AI for social good, privacy-preserving mechanisms must integrate fairness by collaborating with stakeholders, including healthcare institutions and policymakers. Addressing computational and data heterogeneity while maintaining equitable performance is key to preventing societal inequalities. Our research explores privacy-fairness trade-offs in FL, developing frameworks that align with global fairness initiatives like the United Nations Sustainable Development Goals (SDGs) and the Leave No One Behind (LNOB) Principle (Bentaleb and Abouchabaka 2024).

Prior work in federated learning often treats privacy and fairness separately, with privacy studies focusing on differential privacy and encryption under idealized IID settings while ignoring fairness metrics and realistic adversarial threats (Truex et al. 2020; Wang et al. 2020; Zhang et al. 2022; Liu et al. 2022). Conversely, fairness-aware approaches assess equity under controlled conditions but omit privacy-preserving mechanisms and attack analyses (Li, Sanjabi, and Smith 2019; Mohri, Sivek, and Suresh 2019). Most evaluations also rely on limited benchmarks or small-scale simulations, leaving open how DP, HE, and SMC interact with fairness optimizers in heterogeneous and practical domains.

The goal of this work is to investigate the complex interactions between privacy preservation and fairness in federated learning across both benchmark and real-world scenarios in healthcare and finance. We assess the impact of Differential Privacy (DP), Homomorphic Encryption (HE), and Secure Multi-Party Computation (SMC) on client-level fairness metrics and evaluate whether fairness-aware optimization strategies introduce new privacy vulnerabilities. Our empirical study uses four representative datasets: MNIST and Fashion-MNIST for controlled bench-

marks, Alzheimer's Disease MRI scans for medical applications, and a large-scale credit card fraud detection dataset for financial analysis. By conducting experiments under both IID and non-IID client distributions, we aim to chart the multi-dimensional trade-offs that practitioners face when deploying FL in sensitive domains.

By quantifying the trade-offs between privacy and fairness, we deliver actionable guidance for designing federated learning frameworks that uphold strong confidentiality guarantees while ensuring equitable outcomes across diverse clients. Our results offer domain-specific recommendations under regulations such as HIPAA and GDPR for medical use cases and the EU AI Act for financial services. We further propose a set of best practices and parameter guidelines that practitioners can adopt to balance differential privacy budgets, encryption parameters, and fairness weights. This work advances responsible AI by providing empirical evidence and practical strategies that mitigate fairness disparities introduced by privacy mechanisms, enabling trustworthy FL deployment in real-world settings.

This work presents the following **key contributions**:

- We present the first unified, large-scale empirical evaluation of DP, HE, and SMC under fairness-aware optimizers on both IID and non-IID client distributions across benchmark and real-world datasets.
- We systematically quantify trade-offs between privacy-preserving techniques and client fairness under simulated attacks, offering detailed insights into their behavior in heterogeneous federated learning.
- Our analysis shows that HE and SMC achieve a superior balance between privacy and fairness compared to DP, particularly in non-IID scenarios where DP noise disproportionately degrades utility for underrepresented clients, while HE incurs higher computational overhead.
- We uncover counterintuitive interactions between privacy mechanisms and fairness optimizers, demonstrating that DP can erode fairness and that fairness-aware updates can weaken privacy defenses.
- We propose a comprehensive validation framework integrating privacy-fairness schemes with adversarial threat models, offering practical guidelines for robust federated learning in healthcare and finance applications.

Related Work

Privacy-Preserving FL Various studies propose privacy-preserving (PP) techniques for FL, often combining DP and SMC to balance accuracy and privacy, as SMC risks inference attacks and DP reduces accuracy (Truex et al. 2019). Methods like Confined Gradient Descent and blockchain-based optimization reduce cryptographic overhead (Zhang et al. 2024). Cryptographic tools, Trusted Execution Environments (TEEs), and decentralized frameworks are critical, especially in healthcare FL. HE defends against reconstruction attacks (Zhang et al. 2022), while TEE-based FL secures clients and servers (Yazdinejad, Dehghantanha, and Srivastava 2023). Decentralized FL enhances robustness and communication efficiency (Tian et al. 2023). PP FL models support Alzheimer's and COVID-19 detection by safeguarding sensitive data (Li et al. 2022a).

Despite advances, fairness impacts remain underexplored, and standardized privacy metrics are lacking. Theoretical frameworks dominate, with limited real-world benchmarking, restricting FL applicability.

Fairness-Aware FL Existing studies address fairness in FL by ensuring balanced model performance across heterogeneous client data. Most focus on group fairness, aiming for equitable outcomes across demographic groups (Mohri, Sivek, and Suresh 2019), but overlook client fairness (see Supplementary Material, Section B), which considers variations in dataset sizes, distributions, and resources. Other approaches emphasize fairness in client contributions and performance using aggregation techniques and constrained optimization, particularly in healthcare (Meerza et al. 2022) and finance (Kamalaruban et al. 2024). However, these primarily maintain fairness at the group level rather than ensuring equitable benefits for clients.

Despite progress, fairness-aware methods often lack privacy-preserving mechanisms, leaving FL systems vulnerable to privacy risks. Moreover, many assume IID data, while real-world federated settings are predominantly non-IID (Zhao et al. 2018), highlighting the need for rigorous fairness evaluations under privacy constraints.

Critical Tradeoffs in FL Balancing privacy and fairness in FL remains challenging. Privacy-preserving techniques like DP and HE reduce accuracy, disproportionately affecting underrepresented groups (Chen et al. 2023). Conversely, fairness-enhancing methods, such as subgroup-focused optimization, may heighten privacy risks by overfitting minority groups (Zheng et al. 2023). To mitigate these issues, two-stage frameworks enforce fairness (e.g., demographic parity) before applying privacy techniques like local DP (Padala, Damle, and Gujar 2021). Integrated approaches (Corbucci et al. 2024; Pentyala et al. 2022) optimize privacy and fairness during training.

Hence, unlike prior work that treats privacy and fairness separately, our framework jointly quantifies the privacy–fairness–utility trade-off under both IID and non-IID settings.

Experimental Validation Framework

Datasets

To systematically explore the interplay between privacy, fairness, and accuracy in FL, we employ four datasets spanning benchmark tasks, medical imaging, and financial fraud detection. Each dataset is partitioned under both IID and non-IID regimes to capture realistic client heterogeneity.

- **MNIST:** The MNIST dataset (LeCun 1998) consists of 70,000 grayscale images (28×28 resolution) of handwritten digits. Data is normalized to [0,1] and partitioned across $K = 50$ clients. IID partitions distribute data uniformly, while non-IID partitions follow a Dirichlet-based approach (Li et al. 2022b) ($\alpha = 0.5$) to simulate realistic data heterogeneity.

Parameter	Value
Batch size	512
Communication rounds	20
Local training epochs	40
Simulation runs	10
Learning rate	0.1
Learning rate lambda	0.1
Total number of clients	50
Fraction of clients per round	0.1

Table 1: KEY DESIGN PARAMETERS AND THEIR DEFAULT VALUES

- **Fashion-MNIST:** Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017) includes 70,000 grayscale images (28×28 resolution) across 10 clothing categories. Preprocessing and partitioning methods are identical to MNIST, with the same Dirichlet parameter ($\alpha = 0.5$), introducing increased visual complexity to test robustness.
- **Alzheimer's Disease MRI (AD-MRI):** The AD-MRI dataset (Falah.G.Salieh 2023) comprises 4,320 T1-weighted MRI scans labeled Alzheimer's or healthy. After preprocessing (skull-stripping, normalization, resizing to 64×64), SMOTE (Chawla et al. 2002) balances classes. Data is distributed across $K = 50$ hospitals, with IID splits equally balanced, and non-IID splits reflecting specialized (80% Alzheimer's) versus general hospitals (80% healthy).
- **Credit Card Fraud Detection (CCFD):** The CCFD dataset (Dal Pozzolo et al. 2017) contains 284,807 transactions (0.17% fraudulent). After standardization, SMOTE oversamples fraud cases to 5%. Data is distributed among $K = 50$ institutions: IID splits maintain equal fraud rates, whereas non-IID splits vary fraud proportions (10%, 1%, and 0.5%), simulating realistic banking heterogeneity.

Hyperparameters

Hyperparameters were tuned via grid search on a baseline FL model without privacy or fairness enhancements and applied across all FL variants. The key parameters include a learning rate of *0.1*, batch size of *512*, *40* local epochs per client, *20* communication rounds, and *5* participating clients per round. Table 1 summarizes the default values. A CNN with two convolutional layers (32, 64 filters), max-pooling, and a dense layer (1024 units) was used for MNIST, Fashion-MNIST, and MRI, whereas Logistic Regression with L2 regularization (0.01) was used for CCFD.

Threat Model

We assume a single, adaptive adversary $\mathcal{A}$ capable of compromising up to βK clients, where K is the total number of clients and $\beta \in [0, 1]$ is the maximum fraction under adversarial control. In a white-box scenario, $\mathcal{A}$ observes all model updates, gradient exchanges, and protocol metadata but never has direct access to raw client data. This threat model covers two categories of attacks: privacy attacks, which aim to extract sensitive information from model updates, and fairness attacks, which introduce or exacerbate performance disparities among clients.

Privacy Attacks

Membership Inference Attacks These attacks aim to determine whether a specific record was used in training by exploiting model memorization. The adversary collects global model updates and client gradients over multiple rounds, then trains shadow models to approximate client data distributions. By comparing the target model's confidence scores or losses on candidate samples with those of the shadow models, $\mathcal{A}$ can infer membership. The *Membership Inference Attack Success Ratio (MSR)* measures privacy leakage as the fraction of correct guesses minus false positives.

Differential Leakage Attacks These attacks assess how aggregated updates change with the inclusion or removal of a single "canary" record. The adversary alternates between injecting and removing the record locally and observes differences in aggregated gradients. Averaging the ℓ_1-norm differences per parameter across rounds yields the *Differential Leakage Rate (DLR)*, where higher values indicate greater detectability of minor client contributions. DLR complements MSR by focusing on gradient sensitivity rather than model outputs.

Fairness Attacks

Data Poisoning Attacks In data poisoning attacks, the adversary injects malicious samples into 10% of compromised clients' training data (by changing labels randomly) to bias the global model. We quantify impact using *Data Poisoning Attack Accuracy (DPA-A)*, the reduction in clean test accuracy, and *Data Poisoning Attack Accuracy Disparity (DPA-AD)*, the variance in accuracy across clients. High DPA-AD indicates uneven harm, revealing fairness weaknesses.

Backdoor Attacks Backdoor attacks plant triggers in 10% of training samples so that inputs bearing the trigger are misclassified to a target label while clean-data accuracy remains stable. We measure this with *Backdoor Attack Accuracy (BA-A)*, the success rate on triggered inputs, and *Backdoor Attack Accuracy Disparity (BA-AD)*, the client-level variance in BA-A. Elevated BA-AD highlights uneven vulnerability to backdoors across participants.

Metrics

To compare privacy-preserving and fairness-aware FL methods under benign and adversarial conditions, we employ a broad set of quantitative metrics grouped into privacy, fairness, and utility categories. Each metric is computed at the end of training (round T) and averaged over ten independent runs. We report both the mean and the standard deviation.

Privacy Metrics These metrics serve as practical indicators of a federated learning system's vulnerability by translating privacy guarantees into observable outcomes under

realistic adversarial conditions. Rather than relying exclusively on theoretical privacy budgets or cryptographic assumptions, we adopt an empirical perspective that directly measures information leakage through attack simulations. Consequently, our analysis centers on *Membership Inference Attack Success Ratio* (MSR) and *Differential Leakage Rate* (DLR).

- **Membership Inference Attack Success Ratio (MSR):** This metric captures the adversary's ability to distinguish training samples from non-training ones. While various attack strategies exist, we abstract the outcome into a scalar by comparing the model's confidence on held-in versus held-out samples. Let $p_w(y \mid x)$ denote the model's predicted probability for the true label y given input x. We define:

$$\textbf{MSR} = \frac{1}{2N} \sum_{i=1}^{N} \left[p_w(y_i \mid x_i^{\text{in}}) - p_w(y_i \mid x_i^{\text{out}}) \right], \quad (1)$$

  where $\{(x_i^{\text{in}}, y_i)\}$ are samples from the training set and $\{(x_i^{\text{out}}, y_i)\}$ from an equally sized held-out set. A higher MSR indicates greater separation between training and non-training confidences, implying weaker privacy.
- **Differential Leakage Rate (DLR):** This metric evaluates leakage under perturbation-based defenses (e.g., differential privacy) by quantifying how gradients change when a single data point is added or removed. Let $\nabla F_k^+(w)$ and $\nabla F_k^-(w)$ be the gradients from client k with and without the target example. We define:

$$\textbf{DLR} = \frac{1}{d} \left\| \nabla F_k^+(w^{(T)}) - \nabla F_k^-(w^{(T)}) \right\|_1, \quad (2)$$

  where d is the total number of model parameters. A lower DLR indicates stronger protection against differential leakage.

Fairness Metrics These metrics evaluate how evenly model performance is distributed across participating clients, ensuring that privacy or optimization choices do not disproportionately harm any subset of users.

- **Loss Disparity (LD):** To assess how equitably the global model performs across clients, we evaluate the variance of per-client training losses. Let $\mathcal{L}_k$ be the empirical loss on client k's local data under the global model $w^{(T)}$. We define:

$$\textbf{LD} = \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{L}_k - \overline{\mathcal{L}} \right)^2, \quad \overline{\mathcal{L}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_k, \quad (3)$$

  where lower LD values indicate more uniform training loss distribution, promoting fairness.
- **Accuracy Disparity (AD):** This metric captures the variance in client-level model accuracy. Let Acc_k denote the accuracy on client k's held-out data. We define:

$$\textbf{AD} = \frac{1}{K} \sum_{k=1}^{K} \left(\text{Acc}_k - \overline{\text{Acc}} \right)^2, \quad \overline{\text{Acc}} = \frac{1}{K} \sum_{k=1}^{K} \text{Acc}_k, \quad (4)$$

  where lower AD (also referred to as local variance, LV) reflects more consistent model benefit across clients, indicating stronger fairness.

Utility Metrics The utility metrics ensure that privacy or fairness enhancements do not unduly degrade the global model's overall predictive performance.

- **Global Accuracy (Acc):** We assess overall model performance on a held-out test set of size N. Let $\hat{y}_i$ be the predicted label and y_i the true label for example i. We define:

$$\textbf{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i), \quad (5)$$

  where $\mathbf{1}(\cdot)$ is the indicator function. This standard metric verifies whether privacy and fairness defenses preserve overall classification quality.

Federated Learning Schemes

We examine both Fair and Privacy-Preserving (PP) FL schemes to analyze fairness-privacy-accuracy tradeoffs.

Fair FL Schemes

- `q-FedAvg` (Li, Sanjabi, and Smith 2019): Extends FedAvg by giving more weight to clients with higher losses. Fairness is controlled by the parameter q, where $q = 0$ reduces the method to standard FedAvg.
- `q-FedSGD` (Li, Sanjabi, and Smith 2019): Builds on q-FedAvg using FedSGD, where clients send gradient updates after each iteration. The server aggregates them using a q-weighted objective, prioritizing high-loss clients.
- `q-MAML` (Li, Sanjabi, and Smith 2019): Incorporates q-based fairness objectives into Model-Agnostic Meta-Learning (MAML) (Finn, Abbeel, and Levine 2017), adjusting global updates using a q-weighted loss function.
- Agnostic Federated Learning (`AFL`) (Mohri, Sivek, and Suresh 2019): Uses a minimax optimization framework to improve worst-case performance by minimizing the maximum loss over possible data distributions.
- `Ditto` (Li et al. 2021): Simultaneously trains a global model and a personalized model for each client, enabling personalized performance while maintaining generalization across clients.

Privacy-Preserving (PP) FL Schemes

- ***Perturbation-based PP FL:*** `Local Differential Privacy (LDP)` (Truex et al. 2020) adds noise at the client level to protect individual data, though it may reduce model accuracy. `Global Differential Privacy (GDP)` (Wang et al. 2020) injects noise at the server level to mask aggregated updates, offering group-level privacy at the cost of higher noise. `Gradient Masking (GM)` (Boenisch, Sperl, and Böttinger 2021) obfuscates gradients via noise and clipping, similar to LDP, to defend against adversarial reconstruction.
- ***Anonymization-based PP FL:*** K`-Anonymity` ensures each data record is indistinguishable from at least $k-1$ others by grouping gradient updates (Sweeney 2002). L`-Diversity` (Machanavajjhala et al. 2007) enhances

Algorithm 1: q-FedAvg with LDP, GDP, HE, and SMC

Input: K, E, T, q, $1/L$, η, w^0, p_k, $k = 1, \ldots, m$
Initialize: Global model $w^{(0)}$
for $t = 0, \ldots, T-1$ **do**
  Server selects a subset S_t of K devices at random (each device k is chosen with prob. p_k)
  Server sends w^t to all selected devices
  for each selected device $k \in S_t$ **in parallel do**
    Each selected device k updates w^t for E epochs of SGD on F_k with step-size η to obtain $\bar{w}_k^{t+1}$
    Compute local weight update $\Delta w_k^t = L(w^t - \bar{w}_k^{t+1})$
    Compute fairness-adjusted gradient $\Delta_k^t = F_k^q(w^t)\Delta w_k^t$
    Calculate weight $h_k^t = qF_k^{q-1}(w^t)\|\Delta w_k^t\|^2 + LF_k^q(w^t)$
    Apply LDP or GDP:
      LDP: $\tilde{\Delta}_k^t = \Delta_k^t + \mathcal{N}(0, \Delta^2/\epsilon^2)$
      GDP: $\tilde{\Delta}_k^t = \sum_k \Delta_k^t + \mathcal{N}(0, \Delta^2/\epsilon^2)$
    Apply HE: $\mathcal{E}_k(\tilde{\Delta}_k^t) = \text{Enc}(\tilde{\Delta}_k^t)$
    Apply SMC: Split $\tilde{\Delta}_k^t$ into shares $\tilde{\Delta}_k^{tj}$
  end for
  Each selected device k sends $\tilde{\Delta}_k^t$ and h_k^t back to the server
  Server aggregates updates as $w^{t+1} = w^t - \frac{\sum_{k\in S_t}\tilde{\Delta}_k^t}{\sum_{k\in S_t} h_k^t}$
end for
Return final model $w^{(T)}$

K-anonymity by requiring at least l distinct sensitive values per group. T-Closeness (Li, Li, and Venkatasubramanian 2006) ensures that the distribution of sensitive attributes within any group closely matches the global distribution, minimizing data pattern leakage.

- ***Encryption-based PP FL:*** HE (Fang and Qian 2021) allows computations on encrypted data, keeping gradients hidden throughout training. SMC (Liu et al. 2022) enables joint computation across parties without revealing individual data, maintaining confidentiality.

- ***Hybrid PP FL Techniques:*** LDP+HE combines client-side noise and encrypted computation to balance individual privacy and security. GDP+HE applies server-side noise while securing updates via encryption. LDP+SMC integrates client-level noise with secure collaborative computation. GDP+SMC ensures group-level privacy via cryptographic protection. HE+SMC fuses encryption and secure computation for end-to-end PP training.

For the joint analysis of privacy-fairness-accuracy trade-offs, we developed an FL framework integrating q-FedAvg with LDP, GDP, HE, and SMC, as detailed in Algorithm 1.

Simulation Results & Analysis

Effect of Varying Fairness on Privacy Risk

Table 2 reports the most striking privacy-risk variations when adjusting the fairness weight q under IID splits. On MNIST with global DP, loss disparity (LD) climbs from 0.0227 at $q = 0$ to 0.0837 at $q = 10$, differential attack leakage rate (DLR) rises from 0.00267 to 0.00443, and membership-inference attack success ratio (MSR) oscillates near 0.40. Local DP on Fashion-MNIST shows a dramatic drop in LD from 0.447 to 0.00013, while DLR shifts modestly (0.00777→0.00877) and MSR remains in the 0.48–0.51 range. Homomorphic encryption on Fashion-MNIST follows a similar pattern (LD 0.166→0.00004, DLR 0.00220→0.00430, MSR 0.59→0.63). Secure MPC on MNIST yields a reduction in LD (0.0220→0.00483) with only small increases in DLR (0.00059→0.00239) and MSR (0.59→0.60). Under IID MRI, only local DP is sensitive (LD up to 0.00059, DLR to 0.00880, MSR to 0.59), and on IID CCFD all schemes except DP show negligible changes.

Table **??** presents non-IID results. Global DP on MNIST produces LD up to 0.1640 and MSR near 0.55; local DP yields LD in [0.0039,0.0338], DLR around 0.0089, MSR to 0.71. HE on MNIST varies LD from 0.0794 to 0.0028 and MSR from 0.70 to 0.76, while SMC moves LD from 0.0812 to 0.0097. On non-IID Fashion-MNIST, GDP's accuracy is highly sensitive (0.283→0.617→0.352) and LD spans 0.443→0.0997; local DP's LD falls from 0.0338 to 0.0039, DLR to 0.00936, MSR to 0.71. HE and SMC show smaller LD swings and modest DLR/MSR shifts. Non-IID MRI highlights local DP again (LD to 0.00804, MSR to 0.71), with HE also sensitive (LD to 0.00601, MSR to 0.94), while SMC remains essentially flat. On non-IID CCFD, DP's LD stays $\approx 10^{-4}$, DLR around 0.009, and MSR up to 0.66; HE and SMC exhibit no meaningful variation.

These findings indicate that DP-based schemes exhibit the largest sensitivity to fairness tuning, especially in non-IID regimes, often amplifying privacy leakage at intermediate q. In contrast, HE and SMC deliver a more predictable privacy profile when adjusting fairness weights, with only isolated cases of metric variation (Fashion-MNIST, MRI). Designers of DP-protected FL must therefore calibrate q carefully to avoid unintended privacy degradation, whereas cryptographic schemes permit more straightforward fairness adjustments without compromising confidentiality.

Effect of Varying Privacy Levels on Fairness

Table 4 reports how changing the Local Differential Privacy budget ϵ affects fairness metrics under IID splits when $q = 1$. On MNIST, increasing ϵ from 2 to 8 reduces loss disparity (LD) in q-FedAvg from 0.00657 to 0.01260 and in q-FedSGD from 0.00457 to 0.02350, while membership inference attack success ratio (MSR) grows modestly (for example, q-FedAvg's MSR rises from 0.21 to 0.33). Similar trends appear in q-MAML, where LD increases from 0.00464 to 0.01850 and DLR moves from 0.00540 to 0.00723. On Fashion-MNIST, all three algorithms exhibit lower LD at higher ϵ (q-FedAvg: 0.00635 to 0.00220; q-FedSGD: 0.00652 to 0.00268; q-MAML:

MNIST (IID)					FMNIST (IID)				
Scheme	Metric	$q=0$	$q=1$	$q=10$	Scheme	Metric	$q=0$	$q=1$	$q=10$
GDP	LD	0.0227	0.0490	0.0837	GDP	ACC	0.333	0.356	0.316
	DLR	0.00267	0.00449	0.00443		LD	27.25	11.50	5.46
	MSR	0.390	0.350	0.410		DLR	0.00279	0.00439	0.00442
HE	LD	0.0288	0.0168	0.00481	LDP	LD	0.447	0.00220	0.00013
	DLR	0.00225	0.00261	0.00428		DLR	0.00777	0.00825	0.00877
	MSR	0.560	0.520	0.530		MSR	0.480	0.510	0.470
SMC	LD	0.0220	0.0165	0.00483	HE	LD	0.166	0.00496	0.00004
	DLR	0.00059	0.00087	0.00239		DLR	0.00220	0.00252	0.00430
	MSR	0.590	0.600	0.600		MSR	0.590	0.660	0.630
MRI (IID)					**CCFD (IID)**				
Scheme	Metric	$q=0$	$q=1$	$q=10$	Scheme	Metric	$q=0$	$q=1$	$q=10$
LDP	LD	0.00012	0.00044	0.00059	LDP	LD	0.00001	0.00004	0.000001
	DLR	0.00742	0.00842	0.00880		DLR	0.00757	0.00842	0.00881
	MSR	0.520	0.590	0.550		MSR	0.430	0.470	0.480
HE	LD	0.00064	0.00015	0.00074	HE	LD	0.00001	0.00004	0.000001
	DLR	0.00222	0.00256	0.00428		DLR	0.00225	0.00261	0.00428
SMC	LD	0.00086	0.00056	0.00039	SMC	LD	0.00001	0.00004	0.000001
	DLR	0.00058	0.00086	0.00249		DLR	0.00059	0.00087	0.00239

Table 2: q-value sensitivity under IID settings: Only schemes and metrics exhibiting significant variation are shown.

MNIST (Non-IID)					FMNIST (Non-IID)				
Scheme	Metric	$q=0$	$q=1$	$q=10$	Scheme	Metric	$q=0$	$q=1$	$q=10$
GDP	LD	0.0627	0.1640	0.0926	GDP	ACC	0.283	0.617	0.352
	DLR	0.00567	0.00551	0.00533		LD	0.443	0.00866	0.0997
	MSR	0.550	0.480	0.550		DLR	0.00357	0.00573	0.00581
HE	LD	0.0794	0.0911	0.0028	LDP	LD	0.0338	0.0106	0.00392
	DLR	0.00328	0.00399	0.00512		DLR	0.00891	0.00904	0.00936
	MSR	0.700	0.760	0.750		MSR	0.600	0.650	0.710
SMC	LD	0.0812	0.0909	0.00973	HE	LD	0.596	0.00341	0.00269
	DLR	0.00812	0.00095	0.00274		DLR	0.00220	0.00252	0.00430
	MSR	0.750	0.770	0.750		MSR	0.590	0.660	0.630
MRI (Non-IID)					**CCFD (Non-IID)**				
Scheme	Metric	$q=0$	$q=1$	$q=10$	Scheme	Metric	$q=0$	$q=1$	$q=10$
LDP	LD	0.00085	0.0545	0.00804	LDP	LD	0.00017	0.00004	0.000001
	DLR	0.00874	0.00716	0.00972		DLR	0.00933	0.00982	0.00963
	MSR	0.660	0.710	0.700		MSR	0.530	0.600	0.660
HE	LD	0.00196	0.0456	0.00601	HE	LD	0.00017	0.00004	0.00004
	DLR	0.00368	0.00419	0.00596		DLR	0.00328	0.00399	0.00512
	MSR	0.810	0.940	0.850		MSR	0.700	0.760	0.750
SMC	LD	0.0834	0.0524	0.0104	SMC	LD	0.00017	0.00004	0.00004
	DLR	0.00076	0.00098	0.00492		DLR	0.00081	0.00095	0.00274
	MSR	0.750	0.770	0.750		MSR	0.780	0.870	0.810

Table 3: q-value sensitivity under Non-IID settings: Only schemes and metrics exhibiting significant variation are shown.

0.00659 to 0.00288) with small increases in DLR and MSR. These results indicate that larger privacy budgets mitigate unfair noise effects and improve per-client fairness, at the cost of somewhat higher inference risk.

Under Non-IID splits (Table 5), fairness gains require larger ϵ to overcome data heterogeneity. On MNIST Non-IID, q-FedAvg's LD falls from 0.0146 at $\epsilon = 2$ to 0.0732 at $\epsilon = 8$, and q-FedSGD's LD from 0.0131 to 0.0819; both

MNIST (IID)					FMNIST (IID)				
Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
q-FedAvg	LD	0.00657	0.00903	0.01260	q-FedAvg	LD	0.00635	0.00208	0.00220
	DLR	0.00560	0.00718	0.00757		DLR	0.00565	0.00720	0.00763
	MSR	0.210	0.240	0.330		MSR	0.260	0.290	0.320
q-FedSGD	LD	0.00457	0.00544	0.02350	q-FedSGD	LD	0.00652	0.00292	0.00268
	DLR	0.00635	0.00751	0.00755		DLR	0.00633	0.00749	0.00746
	MSR	0.150	0.220	0.340		MSR	0.330	0.350	0.360
q-MAML	LD	0.00464	0.00503	0.01850	q-MAML	LD	0.00659	0.00305	0.00288
	DLR	0.00540	0.00686	0.00723		DLR	0.00534	0.00678	0.00724
	MSR	0.280	0.340	0.350					
MRI (IID)					**CCFD (IID)**				
Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
q-FedAvg	LD	0.000597	0.000845	0.000443	q-FedSGD	LD	0.000237	0.000095	0.000035
	DLR	0.00567	0.00722	0.00755		DLR	0.00642	0.00750	0.00746
q-FedSGD	LD	0.000237	0.000095	0.000035		MSR	0.240	0.300	0.430
	DLR	0.00642	0.00750	0.00746					

Table 4: ϵ-sensitivity under IID settings: Only algorithms and metrics exhibiting significant variation are shown.

MNIST (Non-IID)					FMNIST (Non-IID)				
Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
q-FedAvg	LD	0.0146	0.0307	0.0732	q-FedAvg	LD	0.0173	0.0459	0.0338
	DLR	0.0066	0.0071	0.0086		DLR	0.0077	0.0091	0.0099
	MSR	0.330	0.450	0.480		MSR	0.350	0.510	0.670
	DPA-A	0.167	0.388	0.482					
q-FedSGD	LD	0.0131	0.0285	0.0819	q-FedSGD	LD	0.0147	0.0358	0.0420
	DLR	0.0044	0.0055	0.0095		DLR	0.0083	0.0100	0.0175
	MSR	0.330	0.430	0.560		MSR	0.350	0.500	0.620
	DPA-A	0.123	0.243	0.375					
q-MAML	LD	0.0130	0.0287	0.0824	q-MAML	LD	0.0150	0.0359	0.0411
	DLR	0.0064	0.0080	0.0086		DLR	0.0094	0.0189	0.0328
	MSR	0.290	0.380	0.500					
MRI (Non-IID)					**CCFD (Non-IID)**				
Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	Algorithm	Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
q-FedAvg	LD	0.00060	0.05450	0.00804	q-FedSGD	LD	0.00025	0.00029	0.00002
	DLR	0.00567	0.00722	0.00755		DLR	0.00635	0.00751	0.00755
q-FedSGD	LD	0.00024	0.00009	0.00003		MSR	0.150	0.220	0.340
	DLR	0.00642	0.00750	0.00746					
	MSR	0.240	0.300	0.430					

Table 5: ϵ-sensitivity under Non-IID settings: Only algorithms and metrics exhibiting significant variation are shown.

show DLR rising above 0.008 and MSR exceeding 0.45. On Fashion-MNIST Non-IID, q-FedAvg LD decreases from 0.0173 to 0.0338 and q-FedSGD from 0.0147 to 0.0420, while DLR and MSR also climb. These patterns underscore the need for careful ϵ calibration under skewed data: too small ϵ leads to excessive fairness loss, while too large ϵ can erode privacy.

In contrast, Homomorphic Encryption (HE) fairness under varying polynomial modulus degrees remains largely stable across both IID and Non-IID conditions (see Table 6). Loss disparity and accuracy disparity change by less than 10^{-4} when modulus degrees increase from 4K to 16K. This suggests that cryptographic rather than perturbation-based defenses provide predictable fairness performance independent of parameter tuning. Practitioners should therefore rely on DP budget tuning to manage fairness in noise-based schemes, while leveraging HE's consistent fairness profile when compute and communication costs permit.

Privacy-Fairness Analysis of the SOTA FL Under IID vs. Non-IID Datasets

Performance Analysis of PP FL Algorithms. Table 7 compares IID and non-IID performance across PP FL algorithms. Positive Acc_D values indicate better accuracy on IID datasets, while negative values favor non-IID. MRI shows notable accuracy losses under some algorithms, highlighting sensitivity to data heterogeneity. In contrast, algorithms like K-Anonymity perform well in IID settings (see Table 1 in Section F of Supplementary Material).

Fairness metrics (LV_D, LD_D) improve with smaller differences, with some algorithms maintaining fairness in IID datasets. However, significant fairness losses in non-IID set-

Dataset	Metric	IID			Non-IID		
		4K	8K	16K	4K	8K	16K
MNIST	LD	0.0167	0.0165	0.0166	0.0911	0.0913	0.0877
	DLR	0.00222	0.00124	0.00034	0.00422	0.00332	0.00349
	MSR	0.620	0.610	0.620	0.750	0.740	0.730
FMNIST	LD	0.00496	0.00496	0.00496	0.00341	0.00341	0.00341
	DLR	0.00222	0.00124	0.00034	0.00427	0.00233	0.00147
	MSR	0.690	0.670	0.690	0.810	0.830	0.830
MRI	LD	0.000152	0.000146	0.000142	0.000152	0.000146	0.000142
	DLR	0.00221	0.00141	0.00050	0.00221	0.00141	0.00050
	MSR	0.770	0.760	0.770	0.770	0.760	0.770
	DPA-LV	0.00172	0.00198	0.00198	0.00401	0.00564	0.00576
CCFD	LD	0.0000433	0.0000432	0.0000432	0.0000433	0.0000432	0.0000432
	DLR	0.00225	0.00132	0.00043	0.00225	0.00132	0.00043
	MSR	0.620	0.610	0.620	0.620	0.610	0.620

Table 6: PMD-sensitivity under IID vs. Non-IID for q-FedAvg with HE. Only metrics with significant variation are shown.

tings highlight challenges in mitigating disparities across heterogeneous data. Privacy metrics (MSR_D, DLR_D) tend to favor non-IID datasets when differences are negative, indicating stronger privacy, while higher MSR_D differences suggest better privacy in IID settings but potential trade-offs in fairness and accuracy. Table 7 and absolute metrics in Table S2 of Supplementary Material underscore the need to balance performance differences and absolute values. While differences reveal relative strengths, absolute metrics offer key insights into algorithm suitability across datasets.

Performance Analysis of Fair FL Algorithms. Table 8 summarizes relative performance differences between IID and non-IID datasets for fair FL algorithms, computed as normalized differences relative to IID performance. Positive values indicate better performance in IID settings, while negative values favor non-IID. These differences are relative and do not reflect absolute performance.

For MNIST, q-FedAvg achieves the highest relative accuracy difference (Acc_D) and lowest local variance difference (LV_D), indicating strong adaptability under IID conditions. AFL excels in fairness for non-IID settings with the most negative loss disparity difference (LD_D), while Ditto minimizes MSR success ratio (MSR_D) and differential attack leakage rate (DLR_D), preserving privacy. In FMNIST, Ditto achieves the best relative accuracy and fairness (LV_D), while AFL performs well in fairness and privacy under non-IID settings, showing the lowest LD_D and MSR_D. For MRI, q-MAML shows better accuracy under non-IID conditions, while AFL minimizes LV_D, ensuring fairness. Ditto remains strong in privacy, achieving the lowest MSR_D and DLR_D.

Comparing absolute and relative performance under IID and non-IID datasets (Tables S3 and S4 in Supplementary Material), Ditto consistently excels in fairness and privacy under IID, achieving the lowest local variance, loss disparity, and DLR while maintaining high accuracy for FMNIST and MRI. Under non-IID conditions, AFL performs well in MNIST and MRI, excelling in accuracy, fairness, and privacy. For FMNIST, Ditto remains strong, minimizing DLR and loss disparity. In CCFD, all schemes perform similarly, but Ditto shows the most consistent privacy and fairness, making it reliable for financial fraud detection.

Overall, Ditto shows strong versatility across IID and non-IID scenarios, particularly in privacy and fairness, while AFL adapts well to non-IID conditions, highlighting the need to align algorithms with dataset distributions and trade-offs among accuracy, fairness, and privacy.

Additional Experiments and Discussions. Due to space constraints, further results and analyses are provided in the Supplementary Material. There, Tables S5 and S6 report lattice-attack success ratios for HE and share-reconstruction outcomes for SMC, illustrating how encryption parameters and share thresholds affect security. The Supplementary Material also examines fairness–privacy trade-offs in federated learning for medical imaging and fraud detection in more detail, and discusses regulatory and policy considerations for adaptive privacy frameworks that maintain both compliance and equity across institutions.

Design Guidelines and Best Practices

Drawing on our extensive empirical analysis of DP, HE, SMC and fairness-aware optimizers under both IID and non-IID client distributions, we now provide detailed, actionable design principles to guide the development and deployment of privacy-fair federated learning systems.

Match Privacy Mechanism to Data Heterogeneity

Selecting an appropriate privacy mechanism requires a clear understanding of client data distributions and domain risk tolerances. DP offers formal leakage guarantees by injecting calibrated noise, but our simulations (Table 5) show that in highly skewed, non-IID scenarios this noise can disproportionately degrade minority client accuracy, increasing loss disparity by up to 20 percent. HE preserves exact gradients and maintains fairness metrics (LD, AD) within five percent of the baseline, yet incurs two to three times higher computation cost. SMC strikes a middle ground, offering strong privacy with moderate performance cost, but requires careful threshold tuning to balance fault tolerance and collusion resistance (Table 7). Practitioners should profile data skew using Dirichlet or label-based partition statistics, then choose

PP Algo.	MNIST					FMNIST				
	Acc_D	LV_D	LD_D	MSR_D	DLR_D	Acc_D	LV_D	LD_D	MSR_D	DLR_D
LDP	0.062	-0.005	-0.059	-0.020	-0.004	0.135	-0.028	-0.086	-0.040	-0.002
GDP	0.045	-0.006	-0.040	-0.040	-0.002	0.050	0.022	-0.001	-0.030	-0.001
k-Anonymity	0.237	-0.000	-0.061	0.000	-0.002	-0.107	-0.024	-0.040	-0.040	-0.001
l-Diversity	0.067	-0.007	-0.074	0.000	-0.001	0.254	-0.088	-0.473	-0.030	-0.001
t-Closeness	0.065	-0.007	-0.075	-0.040	-0.001	**0.339**	-0.011	**0.574**	-0.020	-0.001
GM	0.062	-0.005	-0.057	-0.020	-0.004	0.132	-0.021	-0.084	-0.040	-0.002
HE	0.067	**-0.006**	-0.068	-0.010	-0.001	0.257	**-0.089**	-0.470	-0.030	-0.001
SMC	0.067	-0.006	-0.067	-0.040	-0.000	0.252	-0.087	-0.467	-0.040	-0.000
HE + SMC	0.070	-0.006	-0.068	-0.040	-0.001	0.254	-0.087	-0.348	-0.050	-0.001

PP Algo.	MRI					CCFD				
	Acc_D	LV_D	LD_D	MSR_D	DLR_D	Acc_D	LV_D	LD_D	MSR_D	DLR_D
LDP	0.005	-0.000	-0.001	-0.030	-0.004	0.000	-0.001	-0.001	-0.020	-0.002
GDP	-0.247	**-0.107**	NaN	**-0.050**	-0.002	0.000	-0.002	-0.002	-0.030	-0.002
k-Anonymity	0.025	-0.001	**-0.008**	-0.020	-0.002	0.000	-0.001	-0.001	-0.020	-0.001
l-Diversity	0.012	-0.001	-0.002	-0.040	-0.001	0.000	-0.002	-0.002	-0.030	-0.002
t-Closeness	0.011	-0.000	-0.002	-0.020	-0.001	0.000	-0.002	-0.002	-0.030	-0.002
GM	0.005	-0.000	-0.001	-0.030	-0.004	0.000	-0.001	-0.001	-0.020	-0.002
HE	0.018	-0.000	0.000	-0.030	-0.001	0.000	-0.002	-0.002	-0.030	-0.002
SMC	0.013	-0.000	-0.001	-0.040	-0.000	0.000	-0.001	-0.001	-0.020	-0.001
HE + SMC	0.009	-0.000	NaN	-0.010	-0.001	0.000	-0.002	-0.002	-0.030	-0.002

Note: Each metric x_D ($\text{Acc}_D, \text{LV}_D, \text{LD}_D, \text{MSR}_D, \text{DLR}_D$) denotes the performance difference for x (Accuracy, Local Variance, Loss Disparity, MSR, and DLR) between IID and non-IID datasets. Some LD values are undefined (NaN) under GDP due to excessive noise causing gradient divergence.

Table 7: SUMMARY OF PERFORMANCE DIFFERENCES (NON-IID VS. IID) FOR PRIVACY-PRESERVING FL ALGORITHMS

DP for balanced settings where noise impact is minimal. In domains with strict regulatory or equity requirements, such as healthcare or finance, opt for HE or SMC and budget for additional compute and bandwidth. Document expected overheads and negotiate resource allocations with stakeholders before deployment.

Dynamic Fairness Weight Calibration

Fairness-aware optimizers use a tunable parameter, q, to amplify updates from clients with higher loss. Our ablations reveal that fixed q often undercompensates or overshoots fairness targets as model training progresses and client data evolves. To maintain balanced performance, implement a closed-loop calibration loop: first, instrument your FL system to collect per-client metrics (LD, AD, MSR, etc.) at each communication round. Next, define quantitative thresholds for acceptable disparity (for example, LD$<$0.02) and privacy leakage (MSR$<$0.1). At regular intervals e.g. every five rounds, compute these metrics and adjust q by small increments (e.g. increase by 1 if LD exceeds target, decrease by 1 if global accuracy drops below a domain-specific bound). Log every change in q along with the corresponding per-client metric values and in a separate staging environment, run systematic sensitivity tests by varying q over its plausible range and observing the impact on utility and fairness metrics. Use these results to identify a safe operating interval for q that avoids both accuracy collapse and fairness regression. This dynamic approach ensures that the fairness weight can adjust in production without risking unintended biases or performance drops as data distributions shift.

Automated Attack-Driven Calibration

Adversarial evaluation metrics serve as practical guides for fine-tuning defense parameters in federated learning. In our threat model, the membership inference success ratio (MSR) rose sharply when DP's noise budget ϵ exceeded 8. At the same time, differential attack leakage rates (DLR) remained below tolerance only when ϵ remained under 4 (Figure S2 of the Supplementary Material). Lattice attack resistance (LSR) in HE and share reconstruction robustness (SAR) in SMC similarly exhibit nonlinear behavior as cryptographic parameters vary (Tables S5 and S6 of the Supplementary Material). We advise integrating an automated calibration pipeline into your FL framework. First, define target bounds for each privacy and fairness metric, then perform rapid parameter sweeps in an isolated environment to map parameter values to metric outcomes. Use these response surfaces to select the Pareto-optimal configuration that simultaneously satisfies MSR, DLR, LSR, SAR, LD, and AD thresholds. Finally, embed this calibration logic into the deployment

Fair FL Opt.	MNIST					FMNIST				
	Acc_D	LV_D	LD_D	MSR_D	DLR_D	Acc_D	LV_D	LD_D	MSR_D	DLR_D
q-FedAvg	**0.327**	**-0.0531**	-0.074	-0.030	-0.002	0.005	-0.0061	0.002	-0.070	-0.000
q-FedSGD	0.213	-0.0415	-0.065	-0.030	-0.001	-0.027	-0.01019	-0.001	-0.030	-0.000
q-MAML	0.211	-0.0347	-0.092	-0.030	-0.000	0.002	0.002	-0.005	0.000	-0.001
AFL	0.044	0.0084	**-0.712**	-0.010	-0.001	-0.021	-0.066	**-0.934**	**-0.050**	-0.001
Ditto	-0.018	-0.021	0.058	**-0.040**	**-0.002**	**0.177**	**-0.115**	-0.202	-0.030	-0.000

Fair FL Opt.	MRI					CCFD				
	Acc_D	LV_D	LD_D	MSR_D	DLR_D	Acc_D	LV_D	LD_D	MSR_D	DLR_D
q-FedAvg	-0.042	-0.166	-0.052	-0.020	-0.005	0.001	-0.004	-0.010	-0.005	-0.002
q-FedSGD	-0.090	-0.047	-0.014	-0.040	**-0.006**	0.002	-0.005	-0.007	-0.004	-0.001
q-MAML	**-0.169**	-0.067	-0.022	-0.040	0.000	0.001	-0.004	-0.009	-0.003	-0.001
AFL	0.158	**-0.026**	-0.001	0.030	-0.001	0.003	-0.006	**-0.011**	-0.004	-0.001
Ditto	0.104	-0.000	**0.030**	**-0.040**	-0.001	**0.004**	**-0.007**	-0.013	-0.006	-0.002

Table 8: SUMMARY OF PERFORMANCE DIFFERENCES (NON-IID VS. IID) FOR FAIR FL OPTIMIZERS

workflow so that each new FL instance begins with empirically validated settings. This ensures your system maintains the desired balance of privacy guarantees, equitable performance, and utility in diverse real-world conditions.

Future Research Directions

Future research should advance the co-design of algorithms that jointly optimize privacy, fairness, and utility in federated learning. One promising avenue is a privacy–fairness co-optimizer: for example, a two-stage aggregator that first applies fairness-driven client weighting based on loss disparities and then injects adaptive noise or weighted cryptographic clipping to satisfy a global privacy budget. Studies must move beyond static IID and non-IID partitions to personalized and longitudinal settings, leveraging meta-learning and multi-task approaches to accommodate evolving client distributions. Threat models should capture colluding clients, adversaries alternating between inference and poisoning goals, and hybrid campaigns targeting both fairness and privacy. The field also needs standardized benchmarks, datasets, unified partitioning and threat specifications, and open-source toolkits that integrate DP, HE, SMC, fairness-aware optimizers, and attack simulation. Finally, interdisciplinary collaboration with security, ethics, and policy experts is essential to align FL deployments with GDPR, the EU AI Act, HIPAA, and other high-stakes regulations.

Conclusions

In this work, we have presented the first unified, large-scale empirical evaluation of Differential Privacy (DP), Homomorphic Encryption (HE), and Secure Multi-Party Computation (SMC) under fairness-aware optimization in federated learning. By systematically studying these privacy-preserving mechanisms alongside q-FedAvg, q-MAML, Ditto, and AFL under both IID and non-IID client distributions, and by simulating realistic adversarial threats such as membership inference, differential leakage, lattice attacks, share reconstruction, poisoning, and backdoors, we quantified the complex interactions among privacy, fairness, and utility across benchmark (MNIST, Fashion-MNIST) and real-world (Alzheimer's MRI, credit card fraud) datasets.

Our results reveal several **key findings**: (1) DP is highly sensitive to fairness tuning, with the largest variations in privacy leakage and loss disparity, especially at intermediate fairness weights q, necessitating careful calibration; (2) Cryptographic methods offer more stable privacy-fairness tradeoffs, showing minimal sensitivity to fairness tuning; (3) Higher DP budgets improve fairness but increase inference risk, highlighting a clear fairness-privacy tradeoff; (4) Non-IID data amplifies these tensions, requiring fine-grained budget control; (5) HE ensures fairness stability across parameter tuning, maintaining consistent performance regardless of encryption parameters; (6) Ditto shows consistent strength across datasets, especially in FMNIST and MRI; (7) AFL excels in non-IID settings, particularly for MNIST and MRI; (8) DP mechanisms degrade minority client performance under skew, with loss disparity increasing by up to 20%; (9) Encryption parameter tuning affects cryptographic risk more than fairness, allowing HE and SMC to preserve equity while reducing attack surfaces; and (10) Automated calibration pipelines can optimize tradeoffs, enabling systematic tuning of privacy and fairness settings.

Looking forward, our results highlight the need for joint optimization of privacy, encryption, and fairness under realistic threat models. Researchers should develop multi-objective algorithms that anticipate collusion and adaptive attacks. Practitioners in healthcare and finance can apply these insights to choose protocols compliant with HIPAA, GDPR, and the EU AI Act while protecting vulnerable populations. By providing benchmarks, toolkits, and governance templates, we aim to accelerate secure, fair, and trustworthy federated learning deployments.

Acknowledgments

This research is partially supported by the U.S. Army Research Office (ARO) under Award W911NF-24-2-0241, and by the National Science Foundation (NSF) under Awards 2107450, 2330940, and 2106987. Additional support was provided by the Griffis Institute through the Air Force Defense Research Sciences Program and by the Commonwealth Cyber Initiative (CCI) Southwest Virginia (SWVA).

References

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

Bentaleb, A.; and Abouchabaka, J. 2024. A survey of federated learning approach for the Sustainable Development aspect: eLearning. In *E3S Web of Conferences*, volume 477, 00055. EDP Sciences.

Boenisch, F.; Sperl, P.; and Böttinger, K. 2021. Gradient masking and the underestimated robustness threats of differential privacy in deep learning. *arXiv preprint arXiv:2105.07985*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chen, H.; Zhu, T.; Zhang, T.; Zhou, W.; and Yu, P. S. 2023. Privacy and Fairness in Federated Learning: On the Perspective of Tradeoff. *ACM Comput. Surv.*, 56(2).

Corbucci, L.; Heikkila, M. A.; Noguero, D. S.; Monreale, A.; and Kourtellis, N. 2024. PUFFLE: Balancing Privacy, Utility, and Fairness in Federated Learning. *arXiv preprint arXiv:2407.15224*.

Dal Pozzolo, A.; Boracchi, G.; Caelen, O.; Alippi, C.; and Bontempi, G. 2017. Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems*, 29(8): 3784–3797.

Falah.G.Salieh. 2023. Alzheimer MRI Dataset.

Fang, H.; and Qian, Q. 2021. Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet*, 13(4): 94.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Kamalaruban, P.; Pi, Y.; Burrell, S.; Drage, E.; Skalski, P.; Wong, J.; and Sutton, D. 2024. Evaluating Fairness in Transaction Fraud Models: Fairness Metrics, Bias Audits, and Challenges. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 555–563.

LeCun, Y. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Li, J.; Meng, Y.; Ma, L.; Du, S.; Zhu, H.; Pei, Q.; and Shen, X. 2022a. A Federated Learning Based Privacy-Preserving Smart Healthcare System. *IEEE Transactions on Industrial Informatics*, 18(3): 2021–2031.

Li, N.; Li, T.; and Venkatasubramanian, S. 2006. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, 106–115. IEEE.

Li, Q.; Diao, Y.; Chen, Q.; and He, B. 2022b. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, 965–978. IEEE.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and Robust Federated Learning Through Personalization. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 6357–6368. PMLR.

Li, T.; Sanjabi, M.; and Smith, V. 2019. Fair Resource Allocation in Federated Learning. *CoRR*, abs/1905.10497.

Liu, Z.; Guo, J.; Yang, W.; Fan, J.; Lam, K.-Y.; and Zhao, J. 2022. Privacy-preserving aggregation in federated learning: A survey. *IEEE Transactions on Big Data*.

Machanavajjhala, A.; Kifer, D.; Gehrke, J.; and Venkitasubramaniam, M. 2007. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1): 3–es.

Meerza, S. I. A.; Li, Z.; Liu, L.; Zhang, J.; and Liu, J. 2022. Fair and Privacy-Preserving Alzheimer's Disease Diagnosis Based on Spontaneous Speech Analysis via Federated Learning. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1362–1365. IEEE.

Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic federated learning. In *International Conference on Machine Learning*, 4615–4625. PMLR.

Padala, M.; Damle, S.; and Gujar, S. 2021. Federated learning meets fairness and differential privacy. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28*, 692–699. Springer.

Pentyala, S.; Neophytou, N.; Nascimento, A.; De Cock, M.; and Farnadi, G. 2022. PrivFairFL: Privacy-preserving group fairness in federated learning. In *Proceedings of Algorithmic Fairness through the Lens of Causality and Privacy (AFCP2022) - NeurIPS2022 workshop, 2022*.

Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05): 571–588.

Tian, Y.; Wang, S.; Xiong, J.; Bi, R.; Zhou, Z.; and Bhuiyan, M. Z. A. 2023. Robust and privacy-preserving decentralized deep federated learning training: Focusing on digital healthcare applications. *IEEE/ACM Transactions on computational biology and bioinformatics*.

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

Truex, S.; Liu, L.; Chow, K.-H.; Gursoy, M. E.; and Wei, W. 2020. LDP-Fed: Federated learning with local differential privacy. In *Proceedings of the third ACM international workshop on edge systems, analytics and networking*, 61–66.

Wang, H.; Zhao, Q.; Wu, Q.; Chopra, S.; Khaitan, A.; and Wang, H. 2020. Global and local differential privacy for collaborative bandits. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 150–159.

Weng, W.-H.; Deaton, J.; Natarajan, V.; Elsayed, G. F.; and Liu, Y. 2020. Addressing the real-world class imbalance problem in dermatology. In *Machine learning for health*, 415–429. PMLR.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, Z.; Zhang, Y.; Andrew, G.; Choquette, C.; Kairouz, P.; Mcmahan, B.; Rosenstock, J.; and Zhang, Y. 2023. [Industry] Federated Learning of Gboard Language Models with Differential Privacy. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Yazdinejad, A.; Dehghantanha, A.; and Srivastava, G. 2023. Ap2fl: auditable privacy-preserving federated learning framework for electronics in healthcare. *IEEE Transactions on Consumer Electronics*.

Zhang, L.; Xu, J.; Vijayakumar, P.; Sharma, P. K.; and Ghosh, U. 2022. Homomorphic encryption-based privacy-preserving federated learning in iot-enabled healthcare system. *IEEE Transactions on Network Science and Engineering*.

Zhang, Y.; Sun, R.; Shen, L.; Bai, G.; Xue, M.; Meng, M. H.; Li, X.; Ko, R.; and Nepal, S. 2024. Privacy-Preserving and Fairness-Aware Federated Learning for Critical Infrastructure Protection and Resilience. In *Proceedings of the ACM on Web Conference 2024*, 2986–2997. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701719.

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*.

Zheng, M.; Xue, J.; Sheng, Y.; Yang, L.; Lou, Q.; and Jiang, L. 2023. TrojFair: Trojan Fairness Attacks. *arXiv preprint arXiv:2312.10508*.