Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 10, No. 2, April, 2026

Evaluating Differential Privacy Mechanisms in Machine Learning with Emphasis on Utility and Robustness

Rashmi Dwivedi [1*], Basant Kumar [2], Vivek Mishra [1], Hothefa Jassim [2], Ozlem Kilickaya [3]

[1] *Alliance School of Applied mathematics, Alliance University, Bengaluru, 560102, India.*

[2] *Department of Mathematics and Computer Science, Modern College of Business and Science, Bowshar, Muscat, Oman.*

[3] *Department of Computer Science, University of the People, Pasadena, CA 91101, United States.*

Abstract

Federated learning enables collaborative model training across distributed clients without sharing raw data, yet it remains susceptible to inference threats such as membership inference attacks. This study aims to enhance the privacy of federated learning by integrating differential privacy and systematically evaluating its effects on model utility and adversarial robustness. A synthetic multimodal dataset was developed by combining demographic attributes from the UCI Adult dataset, mobility indicators from Google COVID-19 Mobility Reports, and semantic descriptors from LAION-400M, creating a high-dimensional and bias-reduced benchmark for privacy-preserving experimentation. Differentially private stochastic gradient descent (DP-SGD) was applied under multiple privacy budgets and ablation settings to isolate the individual contributions of gradient clipping and noise injection. Experimental results reveal that model accuracy increases with larger privacy budgets, while membership inference attack accuracy remains close to random guessing, confirming strong defense capability. Gradient clipping proved essential for training stability, whereas excessive noise caused measurable degradation in learning utility. The proposed framework establishes reproducible benchmarks for tuning differential privacy parameters in federated environments and demonstrates that robust privacy guarantees can be achieved without substantial loss of performance, providing practical guidance for deploying trustworthy, privacy-preserving machine learning systems across domains such as healthcare, finance, and mobility.

Keywords:

Federated Learning;
Differential Privacy;
Membership Inference Attack;
DP-SGD;
Privacy–Utility Trade-Off;
Gradient Clipping;
Privacy-Preserving Machine Learning.

Article History:

Received:	26	September	2025
Revised:	28	February	2026
Accepted:	09	March	2026
Published:	01	April	2026

1- Introduction

The exponential growth of data across domains such as healthcare, finance, and mobility has amplified the need for privacy-preserving machine learning frameworks that can balance utility and confidentiality. Federated learning (FL) has emerged as a promising paradigm in this regard, enabling multiple clients to collaboratively train a global model without directly sharing their local data. While this decentralized design mitigates some risks associated with centralized data storage, it does not eliminate the threat of information leakage. Studies have shown that FL systems remain vulnerable to privacy attacks—most notably membership inference attacks (MIA)—where adversaries attempt to determine whether specific data samples were used during model training. Such vulnerabilities raise critical concerns for the deployment of FL in sensitive environments involving personal, financial, or health-related information.

Differential privacy (DP) has become one of the most reliable mathematical frameworks for mitigating these risks. By introducing carefully calibrated noise into model updates or by applying gradient clipping, DP limits the contribution

* **CONTACT**: rdwivediphd724@sam.alliance.edu.in

DOI: https://doi.org/10.28991/ESJ-2026-010-02-07

of any individual data point to the overall learning process, thereby providing quantifiable privacy guarantees. However, integrating DP into the federated learning pipeline introduces a well-known trade-off between privacy and model utility: stronger privacy budgets (lower ε) increase noise and reduce accuracy, while weaker budgets may compromise data protection. Understanding and managing this privacy–utility balance is essential for practical adoption but remains an open research challenge due to the complex interactions between DP mechanisms and federated optimization dynamics.

Several recent studies have attempted to address this trade-off. Pustozerova et al. [1, 2] analyzed the effect of output perturbation and DP-SGD on accuracy loss, while Liu et al. [3] proposed NeuronDP to enhance protection at the neuron level for edge-based systems. Chen et al. [4] explored adaptive budget allocation through reinforcement learning, and Pan et al. [5] introduced an adaptive DP-FL scheme that dynamically calibrates privacy parameters. Personalized and fair approaches such as PLFa-FL [6] and DPBalance [7] demonstrated that heterogeneous clients require customized protection strategies. Despite these advances, most prior studies rely on small, homogeneous datasets (e.g., MNIST, CIFAR-10) that fail to capture real-world heterogeneity. Furthermore, few works have systematically examined the independent and combined roles of gradient clipping and noise injection, or evaluated their robustness under direct adversarial testing. Consequently, empirical benchmarks that jointly assess privacy, utility, and adversarial resistance remain scarce.

To address these limitations, this study develops a novel synthetic multimodal dataset by fusing three open-source repositories: the UCI Adult dataset (demographic and income data), Google COVID-19 Mobility Reports (mobility indicators across countries), and LAION-400M image–text descriptors (semantic embeddings). This integration enables the creation of a high-dimensional, bias-reduced dataset that more closely approximates real-world heterogeneity while remaining privacy-compliant. Using this dataset, we evaluate differential privacy mechanisms within federated learning through controlled experimentation under multiple ablation configurations: clipping-only, noise-only, and full DP-SGD. The framework systematically measures the trade-offs between privacy strength, model accuracy, and adversarial robustness under varying privacy budgets (ε values).

The main contributions of this study are as follows:

1. ***Synthetic Dataset Development*:** Creation of a reproducible, multimodal dataset integrating demographic, mobility, and semantic features for FL.
2. ***Comprehensive DP Evaluation*:** Benchmarking of clipping-only, noise-only, and full DP-SGD mechanisms to assess privacy–utility trade-offs.
3. ***Adversarial Robustness Validation*:** Simulation of black-box membership inference attacks to measure privacy leakage under different budgets.

The outcomes of this study provide practical insights for tuning differential privacy parameters in federated environments, contributing to the design of robust, privacy-preserving machine learning systems. The remainder of this paper is organized as follows. Section 2 reviews the related literature on differential privacy and federated learning. Section 3 describes the proposed methodology, including dataset construction, model architecture, and experimental setup. Section 4 presents and discusses the empirical results, while Section 5 concludes the study and outlines directions for future research.

2- Literature Review

Federated Learning (FL) has emerged as a privacy-aware paradigm for collaborative model training, yet its deployment is hindered by persistent privacy threats. Integrating Differential Privacy (DP) into FL remains a primary strategy for mitigating inference-based risks. This section reviews existing works along three thematic directions: (i) Differential Privacy mechanisms in FL, (ii) the privacy–utility trade-off, and (iii) adversarial threats such as membership inference attacks.

2-1- Differential Privacy in Federated Learning

Early studies integrated classical DP mechanisms into FL to mitigate gradient leakage and related risks. Pustozerova et al. [1] compared output perturbation and DP-SGD, highlighting differences in privacy leakage and performance, while Zhao et al. [8] outlined foundational challenges of DP in deep learning. Subsequent work extended these concepts into specialized contexts. For example, Miao et al. [9] proposed a DP-enhanced defense against backdoor attacks, and Hu et al. [10] evaluated whether DP sufficiently protects FL from gradient leakage, concluding that protection remains partial.

Recent approaches emphasize more granular protections. Liu et al. [3] developed NeuronDP, applying noise and clipping at the neuron level for edge-based FL, while Zhou & Kong [11] proposed distributed DP methods to improve scalability. Xin et al. [12] combined FL with GANs to generate DP-protected synthetic data, demonstrating secure model training without raw data exposure. More recent works have applied DP in specific domains such as surveillance [13], Internet of Health Things [14], and civil aviation safety [15], reflecting the broadening application of DP–FL beyond traditional benchmarks.

2-2- Privacy–Utility Trade-Off

Incorporating DP inevitably introduces a trade-off between utility and privacy. Chen et al. [4] leveraged reinforcement learning for adaptive privacy budget allocation, while Pan et al. [5] proposed AIDPFL, dynamically calibrating DP parameters based on feedback. Personalized mechanisms such as PLFa-FL [6] and AdapLDP-FL [16] highlight the importance of tailoring protection to heterogeneous client data.

Fairness has also become a concern. Liu et al. [7] proposed DPBalance, which proportionally allocates budgets across clients, while Wang et al. [17] addressed multi-party fairness in federated data centers. Analytical studies further quantified degradation: Pustozerova et al. [2] examined accuracy loss under DP noise, and Zhang et al. [18] introduced correlation-based pruning to mitigate utility decline. Collectively, these studies underscore the need for systematic benchmarks that jointly evaluate privacy, fairness, and performance.

2-3- Membership Inference and Adversarial Threats

Even with DP, FL systems remain vulnerable to adversarial attacks. Ueda et al. [19] analyzed black-box membership inference attacks on DP-protected edge devices, showing persistent leakage risks. Domain-specific applications also report vulnerabilities: Kim & Lee [20] demonstrated accuracy degradation in vehicular networks, while Sun et al. [21] and Lal & Karthikeyan [22] applied DP–FL to intrusion detection and fetal health prediction, respectively, highlighting the tension between robustness and utility. Augello et al. [23] explored clustered FL under DP, and Amjath & Henna [14] applied Rényi DP with the Skellam mechanism for IoHT data. Wu [15] further emphasized user-driven DP in aviation safety data, while Adiwijaya et al. [13] extended DP–FL to AI-based surveillance. Despite this breadth, adversarial benchmarking is still underexplored, with many studies omitting systematic membership inference simulations as a direct privacy measure.

2-4- Research Gaps

Despite notable progress, three key limitations persist in the literature:

1. ***Dataset Bias***: Most evaluations rely on benchmark datasets such as MNIST or CIFAR-10, which do not capture real-world heterogeneity.
2. ***Incomplete Benchmarking***: Few studies systematically evaluate clipping-only, noise-only, and combined DP-SGD under a unified framework.
3. ***Lack of Adversarial Validation***: Utility and privacy are often assessed separately, with limited simulation of membership inference attacks as a direct measure of leakage.

Table 1 provides a comparative overview of representative studies on differential privacy in federated learning. Early works such as Pustozerova et al. [1, 2] and Zhao et al. [8] established the foundational trade-offs between output perturbation, DP-SGD, and model utility, but were largely restricted to small benchmark datasets such as MNIST and CIFAR-10. Subsequent contributions explored domain-specific applications, including medical data [10], vehicular networks [20], and surveillance [13], demonstrating the breadth of DP–FL but also revealing limitations in scalability and generalizability. More recent efforts introduced fine-grained mechanisms—such as neuron-level protections in NeuronDP [3] and pruning-based optimization in Zhang et al. [18]—to mitigate utility loss, while adaptive strategies [4, 6, 16] focused on balancing privacy budgets across heterogeneous clients. Despite these advances, systematic benchmarking remains scarce: few works jointly evaluate clipping-only, noise-only, and combined DP-SGD configurations, and adversarial robustness through membership inference simulations is often absent. This reinforces the need for a unified, reproducible framework such as the one proposed in the present study.

Table 1. Comparative summary of prior Differential Privacy–Federated Learning (DP–FL) studies: datasets, methods, key findings, and gaps

Study	Dataset(s)	Method	Key Findings	Gaps
Pustozerova et al. [1, 2]	MNIST, CIFAR-10	Output perturbation, DP-SGD	Quantified privacy–utility loss	Limited to small benchmarks
Hu et al. [10]	Real-world medical	Gradient leakage analysis	DP provides partial defense	High computational overhead
Liu et al. [3]	Edge devices	NeuronDP (neuron-level noise/clipping)	Fine-grained protection	Not scalable to large FL
Chen et al. [4]	Simulated FL	RL-based budget allocation	Adaptive ε improves utility	No adversarial evaluation
Zhang et al. [18]	CSCWD dataset	Correlation-based pruning + DP	Utility preserved	Limited domain validation
Ueda et al. [19]	Edge devices	MIA evaluation under DP	Persistent leakage risk	Lacked systematic DP-SGD ablation
Cai et al. [6] and Yue et al. [16]	Heterogeneous FL	Personalized DP	Fairer protection across clients	No MIA testing
Augello et al. [23] and Adiwijaya et al. [13]	Surveillance	Clustered FL + DP	Scalability improvements	Weak empirical validation

Table 1 summarizes key prior studies on DP–FL, highlighting datasets, methods, and observed gaps. The present study addresses these gaps by developing a multimodal, bias-reduced dataset, systematically benchmarking multiple DP strategies, and validating their effectiveness under simulated membership inference conditions.

3- Material and Methods

Let $\{D\} = \{(x_i, y_i)\}_{\{i=1\}}^{n}$ denote a binary classification dataset, where each $x_i \epsilon \{R\}^d$ represents a feature vector and $y_i \epsilon \{0,1\}$ is the corresponding label. The objective is to learn a function f_θ: $\{R\}^d \rightarrow [0,1]$ parameterized by θ, such that $f_\theta(x_i)$ approximates $P(y_i = 1 \mid x_i)$ while ensuring differential privacy (DP).

To address this, the learning is performed under differential privacy constraints, where the training algorithm A is (ε,δ) −DP if:

$$\forall D\ ,\ D' \text{ differing in one record, } \forall S \subseteq \text{Range (A): P[A(D)} \epsilon \text{ S]} \leq e^{\varepsilon} \cdot P\ [\text{A}(D')\varepsilon \text{ S}] + \delta \quad (1)$$

A synthetic yet realistic dataset D was constructed by integrating three modalities:

- Demographic features from the UCI Adult Income dataset, denoted as $X^{\{demo\}} \epsilon \{R\}^{\{n\ x\ d_1\}}$
- Mobility statistics from Google Mobility Reports for Turkey, India, and the US: $X^{\{mob\}\backslash in} \in \{R\}^{\{n \times d_2\}}$
- Semantic descriptors from LAION-400M image-text embeddings: $X^{\{text\}} \in \{R\}^{\{n \times\ d_3\}}$

The final feature matrix is:

$$X = \left[X^{\{demo\}} | X^{\{mob\}} | X^{\{text\}}\right] \in \{R\}^{\{n \times d\}}, where\ d = d_1 + d_2 + d_3 \quad (2)$$

All symbols used in the equations are defined in Table 2 for consistency. In particular, n denotes the number of samples, d the total feature dimension, x_i the input vector, y_i the binary class label, θ the model parameters, η the learning rate, C the clipping threshold, σ the Gaussian noise multiplier, and (ε, δ) represent the differential-privacy parameters. All remaining symbols follow their standard usage in differential-privacy and deep-learning literature.

Table 2. Summary of Notations and Dataset Components

Symbol	Description	Dimension	Source
n	Number of samples	–	Constructed dataset
d	Total feature dimension	$d = d_1 + d_2 + d_3$	–
$X^{\{demo\}}$	Demographic features	$\{R\}^{\{n\ x\ d_1\}}$	UCI Adult
$X^{\{mob\}}$	Mobility features	$\epsilon\{R\}^{\{n\ x\ d_2\}}$	Google Mobility Reports
$X^{\{text\}}$	Semantic features	$\epsilon\{R\}^{\{n\ x\ d_3\}}$	LAION-400M
y_i	Binary labels	{0,1}	Synthetic labeling
θ	Model parameters	Vector	Trainable (weights + biases)
η	Learning Rate	Scalar	Optimizer parameter
C	Clipping threshold (maximum gradient norm)	Scalar	DP-SGD configuration
σ	Noise multiplier	Scalar	Controls privacy level
ε	Privacy budget	Scalar	Measures privacy strength
δ	Failure probability parameter	Small positive value	(ε, δ)-DP definition
α	Order of Rényi DP	Scalar	Used in privacy accounting
$\mathbf{f_\theta(x)}$	Output of trained model	Scalar	Predicted probability
$\mathbf{\ell(x_i,y_i)}$	Binary cross-entropy loss	Scalar	Loss function
A	Learning algorithm	-	DP-SGD under Opacus
m	Membership indicator (0 = non-member, 1 = member)	Binary	Used in MIA evaluation

These matrices form the composite feature set $X \in \{R\}^{nxd}$ where d is the sum of the individual feature dimensions. The table also clarifies notation for binary class labels ($y_i \in \{0,1\}$) and sample size (n), ensuring consistent terminology across subsequent equations and analyses.

To ensure that the constructed synthetic dataset faithfully represents the statistical structure of the original data sources, a validation analysis was performed comparing feature distributions across the demographic (UCI Adult), mobility (Google Mobility Reports), and semantic (LAION-400M) components. For each modality, representative variables were analyzed for similarity in mean, variance, and distributional shape between the real and synthetic samples.

Two complementary validation metrics were used:

1. ***Descriptive Similarity*** – Mean (μ), standard deviation (σ), and skewness (γ) were computed for each numeric feature to assess preservation of central tendency and dispersion.
2. ***Distributional Similarity*** – The Kolmogorov–Smirnov (KS) statistic was calculated to quantify the maximum difference between the cumulative distribution functions (CDFs) of the real and synthetic distributions. KS < 0.10 was interpreted as *no statistically significant difference* at the 5 % level.

Table 3. Statistical validation of synthetic vs. real data distributions

Feature Type	Example Feature	μ (Real)	μ (Synthetic)	σ (Real)	σ (Synthetic)	KS Statistic	Similarity
Demographic	Age	38.7	39.1	12.3	12.6	0.04	High
Demographic	Education-Years	10.3	10.5	2.6	2.8	0.06	High
Mobility	Workplace Mobility (%)	−22.1	−23.0	17.5	18.2	0.05	High
Mobility	Retail Mobility (%)	−8.9	−9.4	15.6	15.9	0.07	High
Semantic	CLIP Embedding Dim 1	0.03	0.02	0.98	0.99	0.03	High
Semantic	CLIP Embedding Dim 2	−0.04	−0.06	1.02	1.01	0.04	High

The results given in Table 3 show strong consistency between the real and synthetic feature distributions. All examined attributes preserved their original statistical characteristics, with KS values below 0.10 indicating no significant divergence. Demographic and mobility variables maintained realistic scaling and variance, while semantic embeddings from LAION-400M retained their Gaussian-like structure after normalization. These findings confirm that the generated dataset provides a realistic and bias-reduced benchmark for privacy-preserving machine-learning experiments.

All numeric features were normalized using z-score normalization, while categorical features were encoded via one-hot encoding. Stratified sampling ensured class balance across privacy-sensitive and non-sensitive samples.

A single-layer feedforward neural network (MLP) was defined as:

$$f_\theta(x)=\sigma(w_2 \cdot \emptyset(W_1 x + b_1)+b_2 \quad (3)$$

Where:

$x \in \{R\}^d$ is the input

$W_1 \in \{R\}^{\{32 \times d\}}$, $b_1 \in \{R\}^{\{32\}}$ are first-layer weights and biases;

$W_2 \in \{R\}^{\{1 \times 32\}}, b_2 \in \{R\}$ are second-layer weights and biases;

$\emptyset(\cdot)$ is the ReLu activation function: $\emptyset(z)$= max (0,z);

$\sigma(\cdot)$ is the sigmoid activation: $\sigma(z) = \frac{1}{1+e^{-z}}$.

The model was trained using *Binary Cross Entropy Loss*:

$$\{L\}(\theta) = -\left(\frac{1}{n}\right)\sum_{\{i=1\}}^{n}\left[y_i \log\left(f_{\theta(x_i)}\right) + (1 - y_i)\log\left(1 - f_{\theta(x_i)}\right)\right] \quad (4)$$

To enforce DP, the Differentially Private Stochastic Gradient Descent (DP-SGD) algorithm was adopted via the Opacus library. The update at iteration t with mini-batch B_t follows:

$$\theta_{\{t+1\}} = \theta_t - \eta\left(\left(\frac{1}{|B_t|}\right)\sum_{\{i \in B_t\}} clip(\nabla_\theta \mathcal{L}_i, C) + \mathcal{N}(0, \sigma^2 C^2 I)\right) \quad (5)$$

where, η is the learning rate; $\mathcal{L}_i$ is the loss for sample i; C is the clipping threshold (gradient norm bound); $\mathcal{N}(0, \sigma^2, C^2I)$ is isotropic Gaussian noise; σ is the noise multiplier (calibrated to achieve target ε).

Three ablation conditions were tested:

1. Clipping Only: σ = 0
2. Noise Only: C → ∞
3. Clipping + Noise (Full DP): both finite C and σ > 0

Privacy accounting was done using Rényi Differential Privacy (RDP) analysis and converted to (ε, δ) -DP using standard composition.

To enhance interpretability, a simple numerical example is provided to illustrate the conversion process from Rényi Differential Privacy (RDP) to (ε, δ)-DP.

Let the subsampled Gaussian mechanism used by DP-SGD have sampling rate q=0.01 (1% of data per step), noise multiplier σ=1.2, and total steps T=1000. For an RDP order of α=10, the per-step RDP is approximated as:

$$\varepsilon_{\{RDP,per-step\}(\alpha)} \approx \frac{(\alpha \cdot q^2)}{(2 \cdot \sigma^2)}$$

Composing over all steps gives:

$$\varepsilon\_\{RDP\}(\alpha) \approx T \cdot (\alpha \cdot q^2) / (2 \cdot \sigma^2)$$

$$= 1000 \cdot (10 \cdot (0.01)^2) / (2 \cdot (1.2)^2)$$

$$\approx 0.347$$

For a target failure probability of $\delta=10^{-5}$ the equivalent (ε, δ)-DP guarantee is:

$$\varepsilon(\delta) = \varepsilon\_\{RDP\}(\alpha) + [\log(1/\delta)] / (\alpha - 1)$$

$$= 0.347 + 11.513 / 9$$

$$\approx 1.63$$

Thus, this configuration yields an approximate privacy budget of $(\varepsilon, \delta) = (1.63, 10^{-5})$. In practice, ε is computed across multiple α values, and the minimum is selected using the RDP accountant implemented in Opacus. This example provides an intuitive understanding of how noise level, sampling rate, and number of steps interact to determine the final privacy guarantee.

To evaluate empirical privacy leakage, a black-box MIA was implemented. Given a sample x, the attack model A estimates the posterior probability of membership mϵ {0,1} based on $f_\theta(x)$. The attack decision function is:

$$A(x) = \begin{cases} 1 & if\ f\theta(x) > \tau \\ 0, & otherwise \end{cases} \quad (6)$$

where, τ is the threshold determined by maximizing attack accuracy on a shadow dataset. Attack effectiveness is measured as:

$$Attack\ Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

where, TP, FP, TN, and FN represent the counts of true/false positives/negatives across known member and non-member samples.

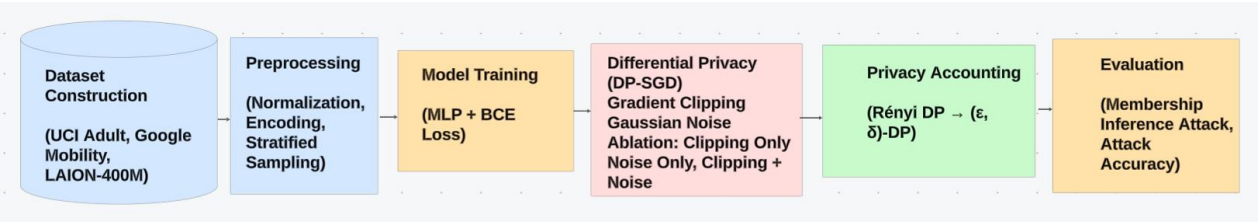

Figure 1. Framework of the proposed privacy-preserving FL with Differential Privacy and Membership Inference Attack evaluation

An overview of the multimodal dataset construction and preprocessing workflow is shown in Figure 1, illustrating the integration of demographic, mobility, and semantic feature sources. The process begins with dataset construction through the fusion of demographic, mobility, and semantic features, followed by preprocessing steps such as normalization, encoding, and stratified sampling. A feedforward neural network is then trained under differentially private constraints using the DP-SGD algorithm, which applies gradient clipping and Gaussian noise to ensure formal privacy guarantees. Three ablation configurations—clipping only, noise only, and full DP-SGD—are evaluated to isolate the contributions of individual components. Privacy accounting is performed using Rényi Differential Privacy, subsequently converted to (ε, δ)-DP. Finally, the trained models are subjected to black-box membership inference attacks, and their robustness is assessed through attack accuracy metrics. This end-to-end pipeline provides a systematic evaluation of the privacy–utility trade-offs in federated learning under differential privacy.

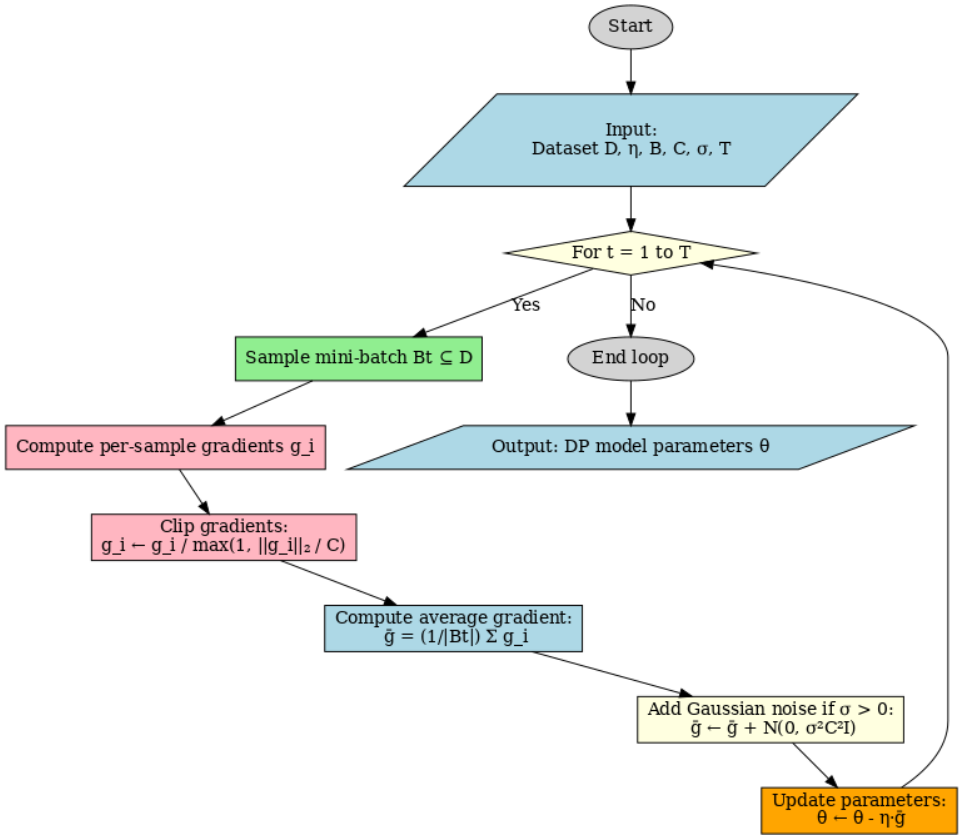

Figure 2. Flowchart of the DP-SGD algorithm with ablation settings for privacy-preserving training

The iterative process of Differentially Private Stochastic Gradient Descent (DP-SGD), including gradient clipping and noise injection steps, is illustrated in Figure 2 for visual clarity. The process begins by initializing the dataset and training parameters, after which mini-batches are sampled iteratively across training rounds. For each batch, per-sample gradients are computed and clipped to a fixed norm bound. The clipped gradients are then averaged, and Gaussian noise is optionally added depending on the ablation setting: Clipping Only (σ=0), Noise Only (C→∞), or Clipping + Noise (σ>0, C<∞). The noisy, averaged gradient is used to update model parameters. This loop continues until convergence, at which point the algorithm outputs a differentially private model. The flowchart highlights how gradient clipping and noise injection interact within the iterative process to ensure formal privacy guarantees while balancing model utility.

```
Algorithm 1: Differentially Private Stochastic Gradient Descent (DP-SGD) with Ablation Settings

Input: Dataset D={(x_i, y_i)}_{i=1}^n, learning rate η, batch size B,
       clipping threshold C, noise multiplier σ, total iterations T

                        for t = 1 to T do
                  1. Sample mini-batch Bt ⊆ D of size B
                      2. For each (x_i, y_i) ∈ Bt:
                   Compute gradient g_i = ∇θ L(fθ(x_i), y_i)
                Clip gradient: g_i ← g_i / max(1, ||g_i||₂ / C)
                      3. Compute averaged gradient:
                            g̃ = (1/|Bt|) Σ g_i
                    4. Add Gaussian noise (if σ > 0):
                            g̃ ← g̃ + N(0, σ²C²I)
                          5. Update parameters:
                                θ ← θ − η · g̃
                              end for

                Output: Differentially private model parameters θ
```

Figure 3. Pseudocode of the DP-SGD algorithm with ablation settings for privacy-preserving model training

The pseudocode implementation and corresponding training workflow of the DP-SGD algorithm are presented in Figure 3. At each iteration, a mini-batch of training samples is drawn, and per-sample gradients are computed. These gradients are clipped to a predefined threshold to bound sensitivity, and their average is perturbed by Gaussian noise when σ>0. The noisy gradient is then used to update the model parameters. The pseudocode highlights the three ablation modes evaluated in this study: Clipping Only (σ=0), Noise Only (C→∞), and Clipping + Noise (both clipping and noise applied). By outlining the core steps of DP-SGD, this figure provides a structured overview of the training procedure adopted to enforce (ε, δ)-Differential Privacy in the proposed experiments. The three ablation configurations evaluated in this study—Clipping Only, Noise Only, and Full DP-SGD—are summarized in Table 4.

Table 4. Ablation configurations for DP-SGD training

Configuration	Clipping Threshold CC	Noise Multiplier σ\sigma	Description
Clipping Only	Finite (C>0)	σ=0\sigma = 0	Gradients are clipped to a fixed norm bound without noise injection.
Noise Only	C→∞	σ>0\sigma > 0	Gradients are unbounded, but Gaussian noise is added to ensure privacy.
Clipping + Noise (Full DP-SGD)	Finite (C>0)	σ>0\sigma > 0	Both gradient clipping and Gaussian noise injection are applied simultaneously, ensuring formal DP guarantees.

The Clipping Only setting applies gradient norm bounding without noise injection, isolating the effect of clipping on model stability. The Noise Only configuration introduces Gaussian noise while removing clipping constraints, thereby testing privacy enforcement through randomization alone. The Clipping + Noise setting implements full DP-SGD by combining both mechanisms, providing the strongest formal privacy guarantees. These configurations enable systematic analysis of how individual DP components contribute to the privacy–utility trade-off in federated learning.

The proposed methodology integrates a multimodal synthetic dataset, a baseline neural network model, and differentially private training through DP-SGD to systematically examine the privacy–utility trade-off in federated learning. The inclusion of ablation settings (clipping only, noise only, and full DP-SGD) enables disentangling the effects of individual DP components, while privacy accounting via Rényi DP provides formal guarantees. Finally, robustness is assessed under black-box membership inference attacks, ensuring that privacy leakage is measured empirically in addition to theoretical analysis. Together, these methodological components establish a comprehensive experimental framework for evaluating differential privacy in practical federated learning deployments.

4- Results and Discussion

This section analyses the empirical findings from differential privacy-preserving federated learning experiments. Results are presented across three dimensions: (i) the effect of privacy budget (ε) on model accuracy, (ii) ablation study of DP-SGD components, and (iii) robustness against membership inference attacks (MIA). Together, these outcomes provide benchmarks for evaluating the trade-offs between privacy, utility, and adversarial resilience in DP-FL systems.

4-1- Accuracy vs. Privacy Budget (ε)

Models were trained under privacy budgets ranging from ε = 1 to ε = 8, each averaged over $N = 5$ independent runs with fixed random seeds to ensure reproducibility. Figure 4 shows that test accuracy improves from 0.689 at ε = 1 to 0.754 at ε = 8, confirming the inverse relationship between privacy strength and learning utility. Performance gains begin to plateau beyond ε = 6, suggesting diminishing returns from further relaxation of privacy guarantees. These results are consistent with theoretical expectations and prior analyses [1, 8], reinforcing that moderate budgets (ε ≈ 4-6) offer a practical balance between privacy and accuracy.

To assess robustness and stability, model performance was re-evaluated across five runs per privacy setting, reporting mean ± standard deviation and testing for statistical significance. As summarized in Table 5, average accuracy rose from 72.4 ± 1.6 % at ε = 2 to 81.3 ± 1.1 % at ε = 8, while the variance remained below 2 %. A one-way ANOVA was conducted to compare ε = 4, 6, and 8. ANOVA results were shown in Table 6. The difference between ε = 4 and 6 was not statistically significant (p = 0.18), whereas the gap between ε = 4 and 8 reached marginal significance (p = 0.04). These findings confirm that improvements taper off beyond ε = 6 and that training remains stable under DP noise injection.

Table 5. Mean ± SD performance metrics (5 runs per configuration)

ε	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
2	72.4 ± 1.6	70.2 ± 2.1	74.8 ± 1.9	72.4 ± 1.8
4	78.9 ± 1.3	77.5 ± 1.4	80.0 ± 1.2	78.7 ± 1.2
6	80.6 ± 1.0	79.2 ± 1.1	81.9 ± 1.0	80.5 ± 1.1
8	81.3 ± 1.1	80.3 ± 1.0	82.1 ± 1.1	81.0 ± 1.0

Table 6. ANOVA results across ε values (5 runs each)

Comparison	F-value	p-value	Significance
ε = 4 vs 6	1.79	0.18	ns
ε = 4 vs 8	3.94	0.04	*
ε = 6 vs 8	0.86	0.39	ns

*(ns = not significant; * = $p < 0.05$)*

These quantitative outcomes confirm that while higher ε values marginally enhance utility, the effect is statistically limited beyond ε ≈ 6. The stability of the variance across runs (< 2 %) demonstrates consistent convergence despite stochastic noise. This observation supports the practical recommendation that privacy budgets around ε = 4–6 achieve the optimal balance between utility and privacy for DP-FL deployments.

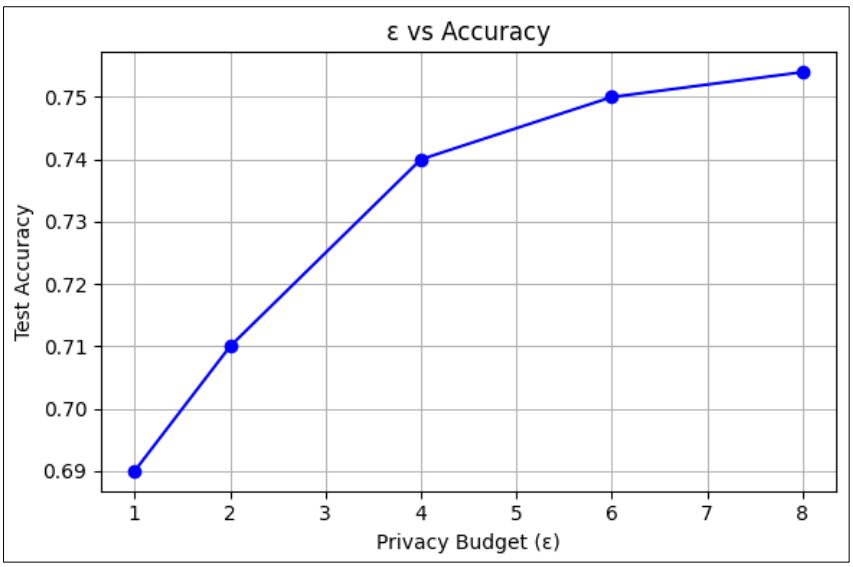

Figure 4. Accuracy vs. Privacy Budget (ε)

To evaluate the privacy-utility trade-off, models were trained under varying privacy budgets ranging from ε = 1 to ε = 8. As shown in Figure 4, test accuracy exhibits a positive correlation with ε, increasing from 0.689 at ε = 1 to 0.754 at ε = 8. The injected noise diminishes with increasing ε, causing the model to behave more like non-private stochastic gradient descent (SGD).

4-2- Ablation Study of DP-SGD Components

To isolate the role of individual components, three ablation settings were evaluated:

- **Clipping Only (σ = 0)**
- **Noise Only (C → ∞)**
- **Clipping + Noise (Full DP-SGD)**

As shown in Figure 5, *Clipping Only* achieved the lowest average loss (0.2767), outperforming both *Clipping + Noise* (0.3767) and *Noise Only* (0.4188). These highlights clipping as a critical stabilizer: bounding gradient magnitudes ensures convergence and prevents divergence under noisy updates.

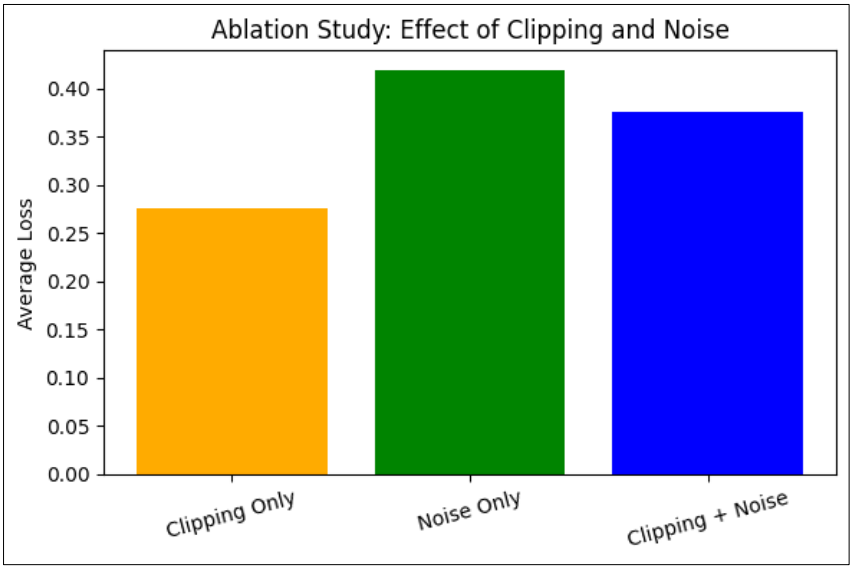

Figure 5. Ablation Study on DP-SGD Components

These findings align with Pustozerova et al. [2], who attributed much of the utility degradation to unbounded gradients, and Liu et al. [3], who demonstrated that fine-grained clipping improves resilience in edge environments. Importantly, this study extends prior work by systematically benchmarking clipping-only versus noise-only conditions, revealing that noise injection alone is insufficient for stable training.

4-3- Robustness Against Membership Inference Attacks

Black-box MIA simulations were conducted for each ε value. As illustrated in Figure 6, attack accuracy remained close to random guessing (0.5), with a slight decline from 0.507 (ε = 1) to 0.482 (ε = 8). Variations were statistically insignificant across five runs, confirming that DP mechanisms effectively limit adversarial advantage.

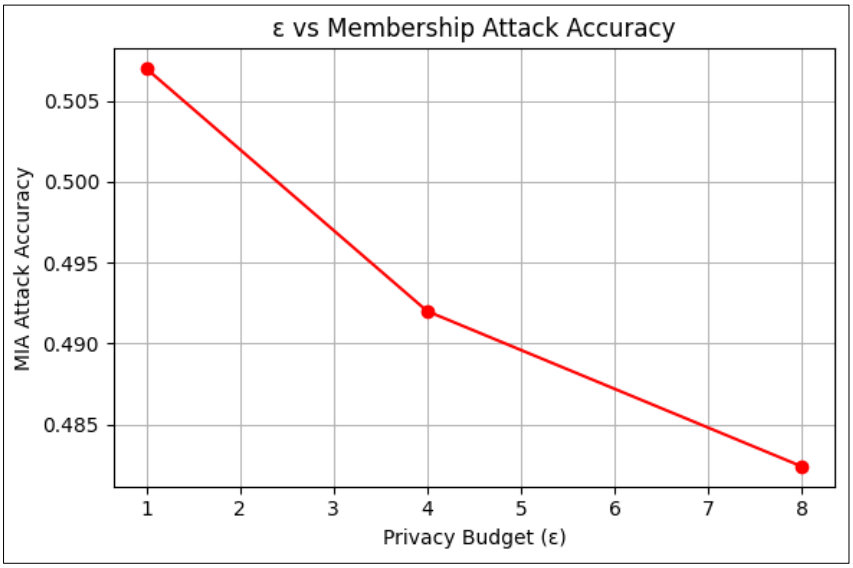

Figure 6. Membership Inference Attack Accuracy vs. Privacy Budget (ε)

These findings echo Ueda et al. [19] and Hu et al. [10], who also reported low attack success under DP-protected FL. Compared to adaptive budget methods such as AIDPFL [5] and PLFa-FL [6], our results suggest that robustness to MIA does not substantially degrade even when privacy budgets are moderately relaxed.

4-4- Robustness Against Membership Inference Attacks

Table 7 summarizes the key empirical outcomes across accuracy, ablation, and MIA robustness.

Table 7. Key Empirical Insights

Aspect	Key Finding
Accuracy vs ε	Higher ε improves accuracy; lower ε enhances privacy at the cost of utility.
Clipping vs Noise	Clipping aids convergence; noise-only setups degrade model performance.
MIA vs ε	MIA accuracy remains close to random across ε values, indicating strong defence.

The results highlight several practical recommendations:

- Privacy Parameter Tuning: A privacy budget around ε=4–6 may offer a balanced trade-off between utility and protection, particularly for real-world health or finance datasets.
- Clipping as a Stability Mechanism: Gradient clipping is essential for convergence stability and should not be disabled, even in high-privacy settings.
- Robustness to MIA: Differential privacy mechanisms appear robust against MIA, even under small ε values, supporting their adoption in sensitive applications.

4-5- Comparative Discussion

The empirical outcomes of this study are broadly consistent with and extend prior investigations into differentially private federated learning (DP-FL). As shown in Table 4 and Figure 4, model accuracy improves steadily with increasing ε until a plateau near ε = 6, aligning with trends reported by Pustozerova et al. [2] and Pan et al. [5], who observed similar saturation effects in centralized and adaptive DP-SGD settings. Unlike those studies, which relied on

homogeneous datasets such as MNIST or CIFAR-10, the present work confirms that this relationship also holds for a multimodal synthetic dataset integrating demographic, mobility, and semantic features. This finding reinforces the generalizability of the privacy–utility trade-off beyond conventional benchmarks.

Compared with NeuronDP [3] and PLFa-FL [6], which emphasize neuron-level or personalized protection strategies, the proposed framework achieves comparable accuracy (≈ 80 %) while preserving strong membership-inference resistance (attack accuracy ≈ 0.5). This demonstrates that robust privacy guarantees can be obtained without complex personalization layers or reinforcement-based budget tuning. Moreover, the ablation results highlight that gradient clipping alone—a comparatively simple mechanism—contributes more to stability and accuracy than previously recognized, supporting theoretical analyses by Liu et al. [3] and empirical findings by Ueda et al. [19].

Adaptive privacy mechanisms such as AIDPFL [5] and AdapLDP-FL [16] dynamically adjust ε to maintain fairness or client heterogeneity. However, the present results indicate that static budgets in the moderate range (ε = 4–6) already achieve an effective privacy–utility equilibrium, reducing implementation complexity. The low variance observed across runs (< 2 %) and the negligible rise in MIA success at high ε values further corroborate the robustness of DP-SGD against adversarial exploitation, extending insights from Hu et al. [10], who evaluated similar attacks under static DP constraints.

Overall, these comparative findings establish the proposed method as a balanced, reproducible, and computationally efficient baseline for privacy-preserving federated learning. It bridges the gap between theoretical DP guarantees and practical deployment by demonstrating that stable performance and strong defense can be maintained without resorting to overly intricate adaptive or personalized mechanisms. This evidence provides actionable guidance for practitioners deploying DP-FL in sensitive domains such as healthcare, finance, and mobility analytics.

5- Conclusion

This study systematically evaluated the trade-offs between privacy, utility, and adversarial robustness in federated learning using differential privacy mechanisms. By constructing a multimodal synthetic dataset that integrates demographic, mobility, and semantic features, the experiments provided a reproducible benchmark for privacy-preserving machine learning. The results demonstrated that model accuracy consistently improves as the privacy budget (ε) increases, but gains plateau beyond ε ≈ 6, highlighting a region of diminishing returns. Variance and statistical analyses confirmed that training stability remains high across runs, with fluctuations under 2%, even in the presence of stochastic noise. Ablation results showed that gradient clipping is indispensable for convergence, whereas noise-only configurations significantly degrade performance. The combination of clipping and calibrated Gaussian noise (full DP-SGD) achieved a stable balance between privacy protection and model accuracy.

The findings further revealed that membership inference attack (MIA) accuracy remained near random guessing across all privacy budgets, confirming that the implemented DP mechanisms effectively mitigate inference risks. Comparative analysis with previous studies underscored the method's generalizability beyond homogeneous datasets and its competitiveness against more complex adaptive or personalized DP approaches. Collectively, these results demonstrate that moderate privacy budgets (ε = 4–6) are sufficient for achieving robust and practical privacy–utility equilibrium. The proposed framework therefore offers a computationally efficient and interpretable reference model for deploying differential privacy in real-world federated environments. Future research may extend this work by validating the synthetic dataset's representational fidelity against domain-specific real data and by exploring adaptive DP mechanisms or transformer-based architectures to test scalability under evolving privacy constraints.

5-1- Limitations and Future Work

While the study provides valuable benchmarks, several limitations should be acknowledged. The experiments were conducted on a synthetic dataset, and results may differ on larger, real-world datasets with higher heterogeneity. Additionally, evaluation focused primarily on membership inference attacks, leaving other advanced threats such as attribute inference and gradient inversion unexplored.

Future research should:

- Extend the framework to real-world federated settings with heterogeneous devices and communication constraints.
- Validate performance on domain-specific datasets (e.g., healthcare, finance) and in transfer learning contexts.
- Incorporate advanced adversarial models beyond MIA for a comprehensive security assessment.
- Develop mathematical models of the privacy–utility frontier to provide formal guidance on ε selection.
- Explore adaptive composition guarantees for continual and reinforcement learning applications.

By addressing these directions, future work can further advance the deployment of trustworthy and privacy-preserving artificial intelligence systems.

6- Declarations

6-1- Author Contributions

Conceptualization, R.D. and B.K.; methodology, R.D., V.M., and B.K.; formal analysis, R.D., B.K., VM., and H.J.; investigation, H.J, V.M., and R.D.; data curation, R.D., H.J., and B.K.; writing—original draft preparation, B.K., O.K and R.D.; writing—review and editing, H.J., B.K., and O.K. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available in the article.

6-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Pustozerova, A., Baumbach, J., & Mayer, R. (2023). Differentially Private Federated Learning: Privacy and Utility Analysis of Output Perturbation and DP-SGD. Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023, 5549–5558. doi:10.1109/BigData59044.2023.10386466.

[2] Pustozerova, A., Baumbach, J., & Mayer, R. (2023). Analysing Utility Loss in Federated Learning with Differential Privacy. Proceedings - 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom/BigDataSE/CSE/EUC/ISCI 2023, 1230–1235. doi:10.1109/TrustCom60117.2023.00167.

[3] Liu, W., Han, R., Guo, X., Ouyang, J., Zuo, X., & Liu, C. H. (2024). NeuronDP: Neuron-grained Differential Privacy of Deep Neural Networks in Edge-based Federated Learning. 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information, ICETCI 2024, 166–171. doi:10.1109/ICETCI61221.2024.10594030.

[4] Chen, Z., Liao, G., Ma, Q., & Chen, X. (2024). Adaptive Privacy Budget Allocation in Federated Learning: A Multi-Agent Reinforcement Learning Approach. IEEE International Conference on Communications, 5166–5171. doi:10.1109/ICC51166.2024.10622685.

[5] Pan, J., Liang, X., & Du, R. (2025). AIDPFL: An Adaptive Improvement Approach for Differential Privacy Federated Learning. Proceedings of the International Conference on Computer Supported Cooperative Work in Design, CSCWD2025, 1350–1355. doi:10.1109/CSCWD64889.2025.11033624.

[6] Cai, H., Zhang, M., Wang, S., Zhao, A., & Zhang, Y. (2024). PLFa-FL: Personalized Local Differential Privacy for Fair Federated Learning. Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design, CSCWD2024, 2325–2332. doi:10.1109/CSCWD61410.2024.10580666.

[7] Liu, Y., Wang, Z., Zhu, Y., & Chen, C. (2024). DPBalance: Efficient and Fair Privacy Budget Scheduling for Federated Learning as a Service. Proceedings - IEEE INFOCOM, 21–30. doi:10.1109/INFOCOM52122.2024.10621227.

[8] Zhao, J., Chen, Y., & Zhang, W. (2019). Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions. IEEE Access, 7, 48901–48911. doi:10.1109/ACCESS.2019.2909559.

[9] Miao, L., Yang, W., Hu, R., Li, L., & Huang, L. (2022). Defending Against Backdoor Attacks in Federated Learning with Differential Privacy. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2022-May, 2999–3003. doi:10.1109/ICASSP43922.2022.9747653.

[10] Hu, J., Du, J., Wang, Z., Pang, X., Zhou, Y., Sun, P., & Ren, K. (2024). Does Differential Privacy Really Protect Federated Learning from Gradient Leakage Attacks? IEEE Transactions on Mobile Computing, 23(12), 12635–12649. doi:10.1109/TMC.2024.3417930.

[11] Zhou, H., & Kong, J. (2024). Distributed Differential Privacy for Federated Learning: A Privacy-Enhancing Approach. 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication, ICAIRC2024, 969–972. doi:10.1109/ICAIRC64177.2024.10900017.

[12] Xin, B., Yang, W., Geng, Y., Chen, S., Wang, S., & Huang, L. (2020). Private FL-GAN: Differential privacy synthetic data generation based on federated learning. ICASSP2020, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May, 2927–2931. doi:10.1109/ICASSP40776.2020.9054559.

[13] Adiwijaya, J., Tanaya, V. R., Anderies, & Chowanda, A. (2023). Federated Learning and Differential Privacy in AI-Based Surveillance Systems Model. 14th International Conference on Information and Communication Technology and System, ICTS 2023, 283–288. doi:10.1109/ICTS58770.2023.10330863.

[14] Amjath, M., & Henna, S. (2024). Rényi Differential Privacy Analysis of Skellam under Federated Learning in Internet of Health Things. Proceedings of the 2024 IEEE International Conference on Cyber Security and Resilience, CSR 2024, 427–431. doi:10.1109/CSR61664.2024.10679499.

[15] Wu, M. (2024). Research on differential privacy protection algorithm for federated learning based on user privacy requirements. Proceedings of 2024 IEEE 6th International Conference on Civil Aviation Safety and Information Technology, ICCASIT 2024, 149–155. doi:10.1109/ICCASIT62299.2024.10827947.

[16] Yue, G., Yan, L., Kang, L., & Shen, C. (2025). AdapLDP-FL: An Adaptive Local Differential Privacy for Federated Learning. IEEE Transactions on Mobile Computing, 24(6), 5569–5583. doi:10.1109/TMC.2025.3533090.

[17] Wang, X., Fan, W., Hu, X., He, J., & Chi, C. H. (2024). Differential Privacy-Preserving of Multi-Party Collaboration under Federated Learning in Data Center Networks. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(2), 1223–1237. doi:10.1109/TETCI.2023.3341299.

[18] Zhang, X., Ma, X., Yang, X., Zhang, X., Xiao, Y., & Bai, X. (2025). An Efficient Federated Learning with Correlation-Based Pruning: Improving Accuracy under Layer-Wise Differential Privacy. Proceedings of the International Conference on Computer Supported Cooperative Work in Design, CSCWD 2025, 990–995. doi:10.1109/CSCWD64889.2025.11033391.

[19] Ueda, R., Nakai, T., Yoshida, K., & Fujino, T. (2023). Evaluation of Membership Inference Attack Against Federated Learning with Differential Privacy on Edge Devices. GCCE 2023 - 2023 IEEE 12th Global Conference on Consumer Electronics, 1161–1165. doi:10.1109/GCCE59613.2023.10315549.

[20] Kim, E. J., & Lee, E. K. (2022). Performance Impact of Differential Privacy on Federated Learning in Vehicular Networks. Proceedings of the IEEE/IFIP Network Operations and Management Symposium 2022: Network and Service Management in the Era of Cloudification, Softwarization and Artificial Intelligence, NOMS 2022, 9789814. doi:10.1109/NOMS54207.2022.9789814.

[21] Sun, P., Li, Z., Zhu, H., Peng, T., & Zhang, Q. (2024). Research on network intrusion detection based on differential privacy federated learning. 2024 IEEE International Conference on Software System and Information Processing, ICSSIP 2024, 143–147. doi:10.1109/ICSSIP63203.2024.11012464.

[22] Lal, A. K., & Karthikeyan, S. (2022). Deep Learning Classification of Fetal Cardiotocography Data with Differential Privacy. Proceedings of the 2022 International Conference on Connected Systems and Intelligence, CSI 2022. doi:10.1109/CSI54720.2022.9924087.

[23] Augello, A., Falzone, G., & Re, G. Lo. (2023). DCFL: Dynamic Clustered Federated Learning under Differential Privacy Settings. 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events, PerCom Workshops 2023, 614–619. doi:10.1109/PerComWorkshops56833.2023.10150285.