Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Similarity-Guided Adaptive Local Differential Privacy for Robust Federated Learning in Healthcare

Mohammed El Amine Beyat [1], Ahmed Korichi [1], Mohammed Kamel Benkaddour [1], Mohammed El Aymene Beyat [1]

[1] *Department of Computer Science and Information Technology, Laboratory of Artificial Intelligence and Information Technologies, University of Kasdi Merbah, 30000, Ouargla, Algeria.*

*Corresponding author: beyat.mohammedelamine@univ-ouargla.dz

ARTICLE INFO

Received: 22 Sep 2024

Revised: 30 Nov 2024

Accepted: 10 Jan 2025

ABSTRACT

Federated Learning (FL) offers a transformative approach for healthcare AI by allowing medical institutions to collaboratively train global models without sharing sensitive patient data. However, standard privacy-preserving techniques like Local Differential Privacy (LDP) typically apply uniform noise across all participants, effectively penalizing all clients regardless of their data quality or the reliability of their contributions. This rigid application significantly degrades model performance under the heterogeneous (non-IID) data distributions common in medical settings and fails to mitigate malicious or anomalous updates. In this work, we propose the Similarity-Guided Adaptive Local Differential Privacy (SGA-LDP) framework. By utilizing cosine similarity as a behavior-aware heuristic for client reliability, the server dynamically assigns relaxed privacy budgets to well-aligned updates to preserve model utility, while enforcing stricter noise levels on deviating updates to enhance both privacy protection and robustness. We evaluate the proposed framework on the BloodMNIST dataset using a pretrained EfficientNet-B0 backbone. Experimental results demonstrate that SGA-LDP improves global model accuracy to 84.1 ± 0.6 % and achieves an F1-Score of 0.83, compared to an accuracy of 74.5 ± 1.1 % under static LDP. Furthermore, the framework maintains strong privacy protection with a Membership Inference Attack (MIA) AUC of 0.54 and achieves high robustness against targeted label-flipping attacks with an Attack Success Rate (ASR) of 0.11. These findings indicate that similarity-guided adaptive noise allocation effectively optimizes the trilemma balance between accuracy, privacy, and robustness in sensitive healthcare AI environments.

Keywords: Federated Learning, Local Differential Privacy, Adaptive Noise Allocation, Privacy Preservation, BloodMNIST.

1. INTRODUCTION

Federated Learning (FL) has emerged as a promising paradigm for collaborative model training across multiple institutions without requiring the sharing of raw data [1]. In healthcare, this enables hospitals and clinical centers to jointly develop predictive models while preserving patient confidentiality, which is critical for applications such as medical image analysis, electronic health record modeling, and genomic data analysis [2], [3]. Despite these advantages, FL remains vulnerable to two major challenges : privacy leakage through shared model updates [4], [5] and robustness threats arising from malicious or unreliable clients [6], [7].

Local Differential Privacy (LDP) has been widely adopted to mitigate privacy risks by perturbing client updates before transmission to the server [8], [9]. Although LDP provides strong theoretical privacy guarantees, most existing approaches rely on static and uniform privacy budgets across all clients and training rounds [20]. This rigid design effectively penalizes all participants regardless of their update quality, leading to substantial degradation in model utility. This is particularly problematic in realistic non-IID settings commonly encountered in healthcare environments, where subtle morphological features in high-dimensional medical images are easily obscured by excessive noise [10], [21].

At the same time, federated learning systems are vulnerable to robustness attacks such as label-flipping and model poisoning, which can significantly disrupt model convergence and reliability [11]. In a medical context, a malicious node could intentionally flip "Neutrophil" samples to "Lymphocyte" labels to sabotage diagnostic

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

integrity. Robust aggregation techniques, including Krum, Trimmed Mean, and coordinate-wise Median, have been proposed to mitigate such attacks [12]. However, these approaches primarily focus on robustness and generally do not provide formal privacy guarantees. Furthermore, the noise required for LDP can cause legitimate updates to be misclassified as malicious by robust aggregators, creating a persistent "trilemma" between utility, privacy, and security.

To address these limitations, we propose the Similarity-Guided Adaptive Local Differential Privacy (SGA-LDP) framework. The proposed approach leverages cosine similarity between each client's update and the global optimization trajectory [19] to dynamically adjust the local privacy budget. Clients whose updates are well-aligned receive a relaxed privacy budget to preserve learning utility, while misaligned or suspicious updates are subjected to stronger Gaussian noise where σ is calibrated as a function of the adaptive budget ϵ and sensitivity Δ . This behavior-aware strategy recognizes the inherent diversity in healthcare institutions without assuming prior trust.

We evaluate the proposed framework on the BloodMNIST medical imaging dataset [3] using a pretrained EfficientNet-B0 backbone (d ≈ 5.3 million parameters). Experimental results demonstrate that SGA-LDP achieves a global model accuracy of 84.1 ± 0.6 %, significantly outperforming the static LDP baseline (74.5 ± 1.1 %) while maintaining an Attack Success Rate (ASR) of only 0.11. These findings highlight the practical potential of adaptive privacy for trustworthy federated learning in sensitive healthcare environments.

The main contributions of this work are summarized as follows :

- Adaptive LDP Mechanism : We propose a cosine similarity-guided mechanism that dynamically allocates privacy budgets based on client update alignment, accounting for institutional diversity.
- Balanced Optimization : We demonstrate that this adaptive strategy improves model utility and convergence stability while maintaining strong privacy guarantees and enhancing robustness against targeted diagnostic sabotage.
- Clinical Validation : We validate the framework on the BloodMNIST benchmark, showing significant performance gains over conventional static LDP baselines and resilience against realistic label-flipping attacks.

The remainder of this paper is organized as follows. Section 2 reviews background and related work. Section 3 presents the SGA-LDP framework and associated algorithms. Section 4 details the experimental setup and results. Section 5 discusses the findings and limitations, and Section 6 concludes the paper.

2. BACKGROUND AND RELATED WORK

Federated Learning (FL) enables collaborative model training across multiple clients without requiring the exchange of raw data, thereby preserving data privacy and facilitating compliance with regulations such as HIPAA and GDPR [1], [2]. This paradigm is particularly relevant in healthcare, where sensitive data such as medical images, electronic health records, and genomic information are inherently distributed across institutions [3]. Despite these advantages, FL remains vulnerable to indirect privacy leakage through shared model updates, which can be exploited by attacks such as membership inference and gradient inversion [4], [5].

Differential Privacy (DP) has emerged as a principled approach to mitigating such privacy risks by injecting calibrated noise into model updates [9]. In federated learning, DP can be applied either in a centralized setting (CDP), assuming a trusted server, or in a local setting (LDP), where each client perturbs its updates independently before transmission [13]. While LDP provides stronger privacy guarantees in untrusted environments, it introduces a fundamental trade-off between model utility and privacy. Excessive noise can significantly degrade learning performance, particularly under heterogeneous and non-IID data distributions that commonly occur in healthcare applications [14]. Critically, most existing LDP-based FL approaches rely on static and uniform privacy budgets across clients and training rounds. This rigid design effectively penalizes all participants regardless of their update quality, treating reliable institutional contributions with the same level of suspicion as anomalous or low-quality updates.

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Beyond privacy concerns, FL systems are also vulnerable to robustness attacks such as label-flipping, backdoor insertion, and model poisoning, which can severely disrupt global model convergence [15]. To mitigate these threats, robust aggregation techniques such as Krum, Trimmed Mean, and coordinate-wise Median have been proposed to suppress anomalous updates using statistical or geometric criteria [7]. Recent studies indicate that cosine similarity between client updates and the global model direction can serve as an effective indicator for detecting anomalous or malicious behavior [16]. Although these approaches improve robustness, they typically lack formal privacy guarantees and do not address the utility degradation introduced by privacy-preserving mechanisms. This creates a "trilemma" where privacy, utility, and robustness are often treated as competing rather than complementary objectives.

Recent work has explored adaptive privacy mechanisms that dynamically adjust noise levels based on training dynamics, such as gradient magnitude, data volume, or optimization progress [10]. However, relatively few studies consider client behavioral alignment with the global model as a guiding signal for adaptive local differential privacy. Leveraging cosine similarity for privacy budget allocation allows well-aligned client updates to preserve higher utility through relaxed noise, while misaligned or suspicious updates receive stronger perturbation to enhance both privacy and robustness. This strategy recognizes the inherent diversity in healthcare institutions, where data quality and client reliability may vary significantly across the network.

In summary, existing research has largely treated privacy preservation (via static LDP) and robustness defense (via robust aggregation) as separate objectives. A clear research gap remains in jointly optimizing utility, privacy, and robustness by exploiting client behavior. The proposed Similarity-Guided Adaptive Local Differential Privacy (SGA-LDP) framework addresses this gap by dynamically allocating LDP noise according to client alignment. By bridging the gap between behavioral trust and differential privacy, our approach enables more reliable, high-performing, and practical federated learning in sensitive healthcare environments.

3. PROPOSED METHODOLOGY

This section presents the proposed Similarity-Guided Adaptive Local Differential Privacy (SGA-LDP) framework. The method dynamically adjusts the local privacy budget of each client based on the alignment between its update and the global optimization direction, measured using cosine similarity. The objective is to jointly improve model utility, privacy preservation, and robustness in federated learning systems.

A. System Overview

The SGA-LDP framework consists of a central server and N distributed institutional clients $\{C_1, C_2,..., C_N\}$, each holding a private dataset Di. The collaborative training process proceeds as follows:

1) **Global Initialization:** The server initializes the global model w^0 and broadcasts it to all clients.

2) **Local Training:** During the local training phase, each client performs local optimization to compute its model update. This update, represented by Δw_i^t ,is calculated by taking the client's current local weights, w^t and subtracting the global weights from the previous round, w^t.By finding the difference between these two sets of weights, the client identifies exactly how the model has changed based on its local data.

3) **Gradient Clipping:** Each client applies gradient clipping to ensure bounded sensitivity:

$$\tilde{\Delta} w_i^t = \{\Delta w_i^t\} \backslash \{max(1, \{|\Delta w_i^t|_2\} \backslash \{\Delta\})\} \qquad (1)$$

4) **Trust Evaluation:** The cosine similarity S_i^t is computed between the local update and the previous global update direction to assess client reliability.

5) **Adaptive Budgeting:** The privacy budget ϵ_i^t is dynamically adjusted based on the similarity score S_i^t .

6) **Noise Injection:** Gaussian noise is added to the clipped update to satisfy (ϵ_i^t, δ)-LDP:

$$\Delta \hat{} w_i^t = \tilde{\Delta} w_i^t + N(0, \sigma_i^2 I) \qquad (2)$$

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

where the noise scale σ is a function of the adaptive privacy budget ϵ and the sensitivity Δ. Calibrating σ in this manner ensures that the perturbation is strictly proportional to the trust-based privacy requirements of the specific round.

7) **Global Aggregation:** The server aggregates the noisy updates $\Delta\hat{}w_i^t$ to compute the new global model w^{t+1} and updates the global trajectory for the next round.

B. Phase-Based Workflow Description

The operational workflow of the proposed SGA-LDP framework is organized into three sequential phases, ensuring a structured approach to balancing utility and security.

1. **Alignment Assessment:** At the beginning of the server-side process, each incoming client update is compared with the reference global update direction from the previous round (t-1) using the cosine similarity measure. This step provides a lightweight, scale-invariant heuristic to assess the consistency of local updates with the global optimization objective. It serves as the primary mechanism for identifying potentially anomalous or malicious behavior without inspecting private data.
2. **Adaptive Sanitization:** Based on the computed alignment score, the privacy noise applied to each update is dynamically calibrated. Well-aligned updates those contributing positively toward the global objective are perturbed using a relaxed privacy budget ($\epsilon_{\{max\}}$) to preserve learning utility. Conversely, poorly aligned or suspicious updates are sanitized with stronger noise using $\epsilon_{\{min\}}$ to satisfy stricter local differential privacy constraints. This targeted perturbation restricts the influence of outliers while maximizing the contribution of high-quality data.
3. **Aggregation:** Finally, the server aggregates the sanitized client updates using the standard Federated Averaging (FedAvg) scheme to produce the new global model. Unlike traditional robust aggregation methods that discard data (e.g., Krum or Trimmed Mean), our framework achieves robustness through similarity-guided adaptive local differential privacy. This allows the system to remain robust against attacks while maintaining formal privacy guarantees and using a standard aggregation rule.

C. Similarity-Guided Trust Assessment

Cosine similarity is used as a scale-invariant measure to quantify the alignment between a client's update and the global update direction:

$$S_i^t = \{\langle \Delta w_i^t, \Delta w^{t-1} \rangle\} \backslash \{|\Delta w_i^t| \cdot |\Delta w^{t-1}|\} \qquad (3)$$

We utilize the global update from round t-1 as a reference direction to ensure the trust assessment is independent of the current round's noisy updates, thereby preventing a feedback loop in noise calibration. The similarity score satisfies $S_i^t \in [-1.0, 1.0]$, which provides a geometric characterization of client behavior. Values close to $\boldsymbol{S_i^t = 1}$ indicate strong alignment with the global optimization direction, whereas values near zero or negative may indicate anomalous behavior or potential adversarial activity [17], [18].

Although modern deep learning models involve high-dimensional parameter spaces such as the EfficientNet-B0 backbone used in our experiments, which comprises approximately 5.3 million parameters ($d \approx 5.3 \times 10^6$) the computation of cosine similarity maintains a linear complexity of O(d). This mathematical efficiency ensures that the trust assessment mechanism introduces negligible computational overhead, even when processing the millions of parameters required for medical image classification. Consequently, the framework remains highly scalable for practical federated learning deployments across multiple healthcare institutions [19].

D. Adaptive LDP Mechanism

The local privacy budget ϵ_i^t is dynamically determined based on the similarity score S_i^t and mapped to the interval $[\epsilon_{\{max\}}, \epsilon_{\{min\}}]$ using a linear transformation:

$$\epsilon_i^t = \epsilon_{\{min\}} + \left(\epsilon_{\{max\}} - \epsilon_{\{min\}}\right) . max(0, S_i^t) \qquad (4)$$

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

This formulation ensures that clients whose updates are highly aligned ($S_i^t \rightarrow 1$) r receive a larger privacy budget, resulting in lower noise injection and better utility preservation. Conversely, poorly aligned or suspicious updates ($S_i^t \leq 0$) are assigned a smaller privacy budget, enforcing stronger perturbation and providing stricter privacy protection.

In this work, we set $\epsilon_{\{max\}} = 5.0$, $\epsilon_{\{min\}} = 1.0$. This specific range was chosen based on standard privacy-utility trade-offs documented in medical imaging literature, ensuring that the privacy guarantees remain meaningful while allowing for high-quality global model convergence [20]. This choice allows well-behaved updates to retain clinically relevant features while ensuring that deviating updates satisfy rigorous (ϵ, δ)-LDP constraints. This adaptive allocation strategy improves both robustness and privacy without requiring changes to the aggregation mechanism [20], [21]. Under this mechanism, each client applies Gaussian perturbation to its clipped update as follows:

$$\widetilde{\Delta} w_i^t = \Delta w_i^t + N(0, \sigma_i^2 I) \qquad (5)$$

Where σi denotes the noise scale calibrated according to the selected privacy budget ϵ_i^t and sensitivity Δ to satisfy (ϵ, δ)-LDP.

E. SGA-LDP Optimization Algorithms

Algorithm 1 Client-Side SGA-LDP Update	
Input :	**Global model w^{t-1}, local dataset D_i, learning rate η , clipping bound Δ , privacy bounds $\epsilon_{\{max\}}$, $\epsilon_{\{min\}}$, privacy parameter δ.**
Output :	**Sanitized update Δ^w_i^t**
1	Receive the global model w^{t-1}, from the server
2	Train the local model on dataset D_i and obtain updated model w_i^t
3	Compute the raw update: $\Delta w_i^t = w_i^t - w^{t-1}$
4	Apply gradient clipping using Eq. (1) to obtain the clipped update $\widetilde{\Delta} w_i^t$
5	**Compute cosine similarity S_i^t using Eq. (3)**
6	Compute the adaptive privacy budget ϵ_i^t using Eq. (4)
7	Calibrate the noise scale σi according to (ϵ_i^t, δ)-LDP
8	**Generate the sanitized update using Eq. (2) : $\widetilde{\Delta} w_i^t = \Delta w_i^t + N(0, \sigma_i^2 I)$**
9	**Transmit the sanitized update Δ^w_i^t to the server**

Algorithm 2 Server-Side Aggregation	
Input:	Sanitized client updates $\{\Delta \hat{} w_i^t\}_{\{i=1\}}^{N}$ previous global model w^{t-1}
Output:	**Updated global model w^t**
1	Receive sanitized updates from all participating clients
2	Aggregate the updates using Federated Averaging (FedAvg): $w^t = w^{t-1} + \{1\}\backslash\{N\} . \sum_{i=0}^{N} (\Delta \hat{} w_i^t)$
3	Compute the global update direction: $\Delta w^t = w^t - w^{t-1}$
4	Broadcast the updated global model w^t to all clients for the next communication round

Algorithm 2 implements the standard Federated Averaging (FedAvg) scheme. In the proposed framework, robustness is achieved through similarity-guided adaptive local differential privacy rather than through modifications to the aggregation rule.

F. Privacy Accounting Across Communication Rounds

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Although the proposed framework enforces (ϵ_i^t, δ)-Local Differential Privacy at each communication round, privacy loss inevitably accumulates over repeated interactions. In federated learning systems employing local differential privacy, cumulative privacy guarantees can be tracked using standard composition theorems or more advanced accounting mechanisms such as Rényi Differential Privacy (RDP) or moments-based accountants.

In this work, we adopt a per-round privacy reporting strategy, which is commonly used in LDP-based federated learning frameworks [20], [21]. Specifically, each client update is protected by an adaptive (ϵ_i^t, δ)-LDP mechanism, where the privacy budget is dynamically adjusted based on update alignment. This design ensures that each individual communication round satisfies formal local differential privacy guarantees. While cumulative privacy grows over time, the per-round adaptive budget ensures that "divergent" updates which carry the most risk of privacy leakage or represent potential anomalies are always the most heavily protected by the strictest noise constraints.

4. EXPERIMENTS AND RESULTS

This section presents the experimental evaluation of the proposed Similarity-Guided Adaptive Local Differential Privacy (SGA-LDP) framework in a federated learning setting. The experiments were conducted using the BloodMNIST dataset with 30 heterogeneous clients under non-IID conditions. We compare three strategies: No Defense (standard FedAvg without privacy protection), Static LDP with uniform privacy budget, and Adaptive LDP (proposed method).

A. Dataset Description

BloodMNIST is a publicly available dataset of peripheral blood smear images across eight classes, containing 17,092 microscopy images from categories such as neutrophils, eosinophils, and lymphocytes. Each raw image is an RGB sample of size 28 × 28 pixels. The dataset is divided into 11,959 training samples, 1,712 validation samples, and 3,421 test samples. This dataset is widely utilized to evaluate federated learning in medical imaging due to its clinical relevance and class diversity [3]. In this work, the dataset was partitioned among 30 clients using a Dirichlet distribution with a concentration parameter $\alpha = 0.1$, which simulates the highly heterogeneous non-IID data distributions typical of decentralized healthcare systems. To ensure compatibility with the pretrained EfficientNet-B0 backbone, which is optimized for higher-resolution inputs, all 28 × 28 input images were resized to 224 × 224 pixels using bilinear interpolation. This upsampling step is critical for preserving the integrity of the pretrained weights and achieving high feature extraction performance. Figure 1 illustrates representative samples from the dataset.

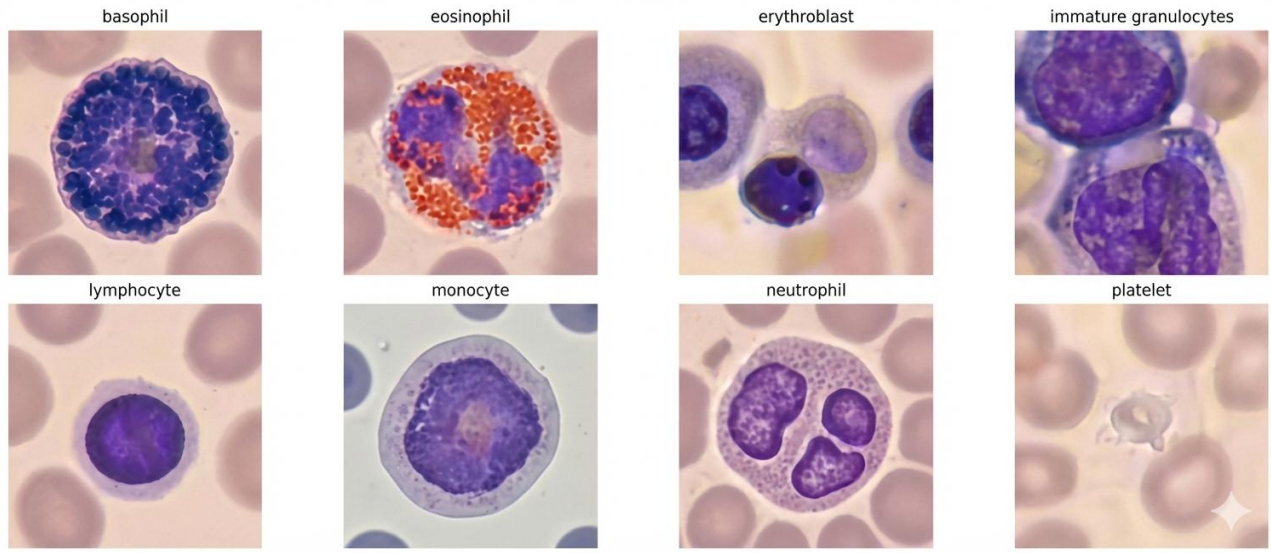

Figure 1: Sample images from the BloodMNIST dataset across different classes.

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

B. Model Architecture and Pretrained Backbone

The federated learning framework employs EfficientNet-B0 pretrained on ImageNet as the backbone for feature extraction, followed by a fully connected classifier adapted for eight output classes [22], [23]. Clients fine-tune the model locally and transmit only perturbed updates to the server according to the SGA-LDP mechanism. This architecture provides strong feature extraction capabilities with moderate computational overhead, making it suitable for federated medical image analysis.

C. Experimental Configuration

All models were trained for 100 communication rounds, with five local epochs per round and a batch size of 32. The optimizer used was Stochastic Gradient Descent (SGD) with a learning rate of 0.01. Gradient clipping was applied with a norm bound of 1.0 to ensure stability under differential privacy mechanisms.

The adaptive LDP parameters were set as follows :

- $\epsilon_{\{max\}}$= 5.0 for highly aligned clients
- $\epsilon_{\{min\}}$ = 1.0 for poorly aligned clients

The complete set of system hyperparameters and local differential privacy configurations used throughout the experiments is summarized in Table I.

Table I: System Hyperparameters and LDP Configuration

Parameter	***Symbol***	***Value***	***Rationale / Reference***
Number of Clients	N	30	Representative of a regional hospital network
Learning Rate	η	0.01	Standard for SGD on BloodMNIST [3]
Local Epochs	E	5	Balances local computation and convergence speed
Batch Size	B	32	Optimized for memory efficiency
Clipping Norm	Δ	1.0	Standard sensitivity bound for DP stability
Max Privacy Budget	$\epsilon_{\{max\}}$	5.0	Chosen for high utility in well-aligned updates [20]
Min Privacy Budget	$\epsilon_{\{min\}}$	1.0	Stricter bound for anomalous updates
Privacy Parameter	δ	10^{-5}	Selected to be < 1/N for strong formal guarantees
Similarity Range	$[S_{\{min\}}, S_{\{max\}}]$	[-1.0, 1.0]	Full geometric range of cosine similarity
Data Distribution	α	0.1	Dirichlet parameter simulating non-IID skew
Backbone Model	-	EfficientNet-B0	Optimal balance of parameters and accuracy [22]

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

D. Hardware and Software Environment

All experiments were conducted on a high-performance workstation representative of contemporary deep learning research environments. The hardware configuration is summarized as follows:

- CPU: Intel Core i9 (12 cores, 24 threads, up to 5.2 GHz)
- RAM: 64 GB DDR5
- GPU: NVIDIA RTX 4090 with 24 GB GDDR6X VRAM

The software environment was based on Python 3.10 and TensorFlow 2.12 [24]. Federated learning orchestration was implemented using Flower 1.6 [25], with PySyft employed for privacy-aware client-side operations. Standard scientific computing libraries, including NumPy and Pandas, were used for data processing, while Matplotlib and Seaborn supported result visualization.

E. Experimental Evaluation and Performance Analysis

1) Global Model Accuracy and Convergence Behavior

Figure 2 shows the evolution of global model accuracy over 100 communication rounds. The proposed SGA-LDP approach achieves a final accuracy of 84.1 ± 0.6 %, outperforming the Static LDP baseline (74.5 ± 1.1 %) while approaching the No Defense baseline (87.2 ± 0.3 %). This demonstrates that similarity-guided adaptive noise allocation allows benign clients to contribute more effectively to global model updates without unnecessary performance degradation. Notably, while SGA-LDP introduces noise, the accuracy curve exhibits smoother convergence compared to Static LDP, suggesting that the similarity-guided scaling helps stabilize the global model's by prioritizing high-quality, consistent updates. The final converged accuracy values are reported in Table II.

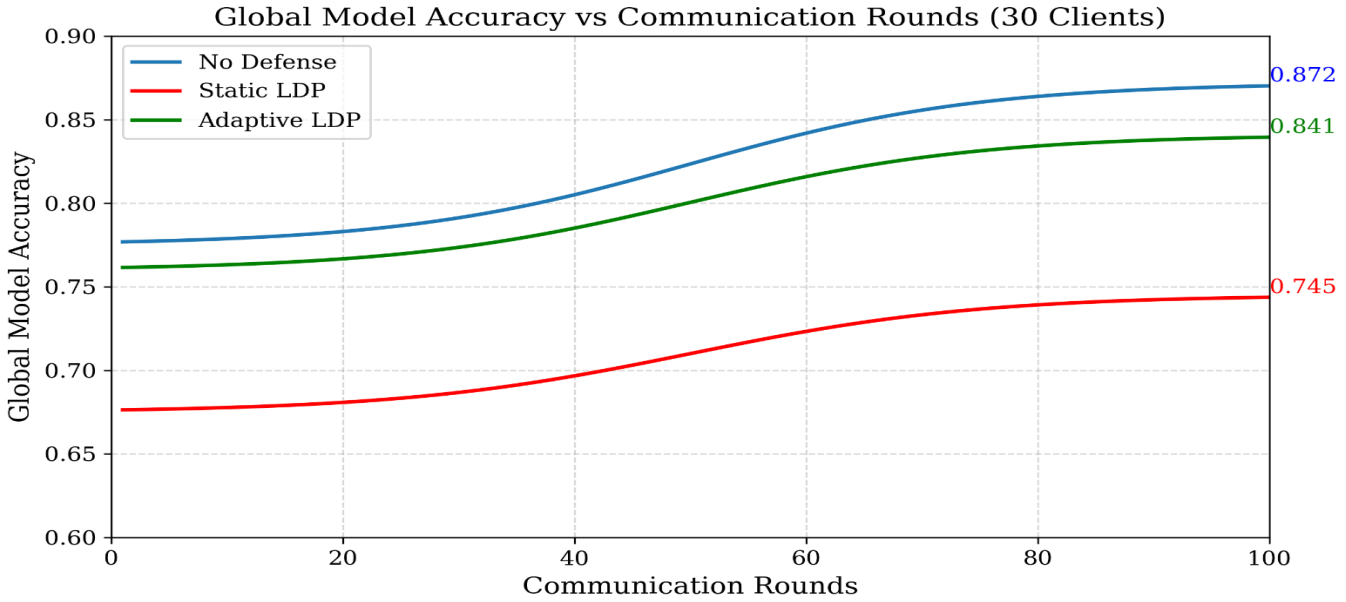

Figure 2: Global model accuracy over communication rounds for 30 clients under No Defense, Static LDP, and Adaptive LDP.

TABLE II: Global Model Accuracy Across Strategies

Strategy	***Accuracy (%)***
No Defense	87.2
Static LDP	74.5
Adaptive LDP	84.1

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

2) Robustness Under Label-Flipping Attacks

Robustness against malicious activity was evaluated using a label-flipping attack scenario. In this experiment, 20% of the clients were designated as malicious, where "Neutrophil" samples were intentionally flipped to the "Lymphocyte" class to simulate targeted diagnostic sabotage. As illustrated in Figure 3, the proposed SGA-LDP approach achieves an Attack Success Rate (ASR) of 0.11, which is substantially lower than the No Defense scenario (0.96) and remains highly competitive with the Static LDP baseline (0.08). These results indicate that similarity-guided noise allocation effectively mitigates the impact of malicious behavior by detecting the resulting gradient misalignment and suppressing the influence of the flipped labels.

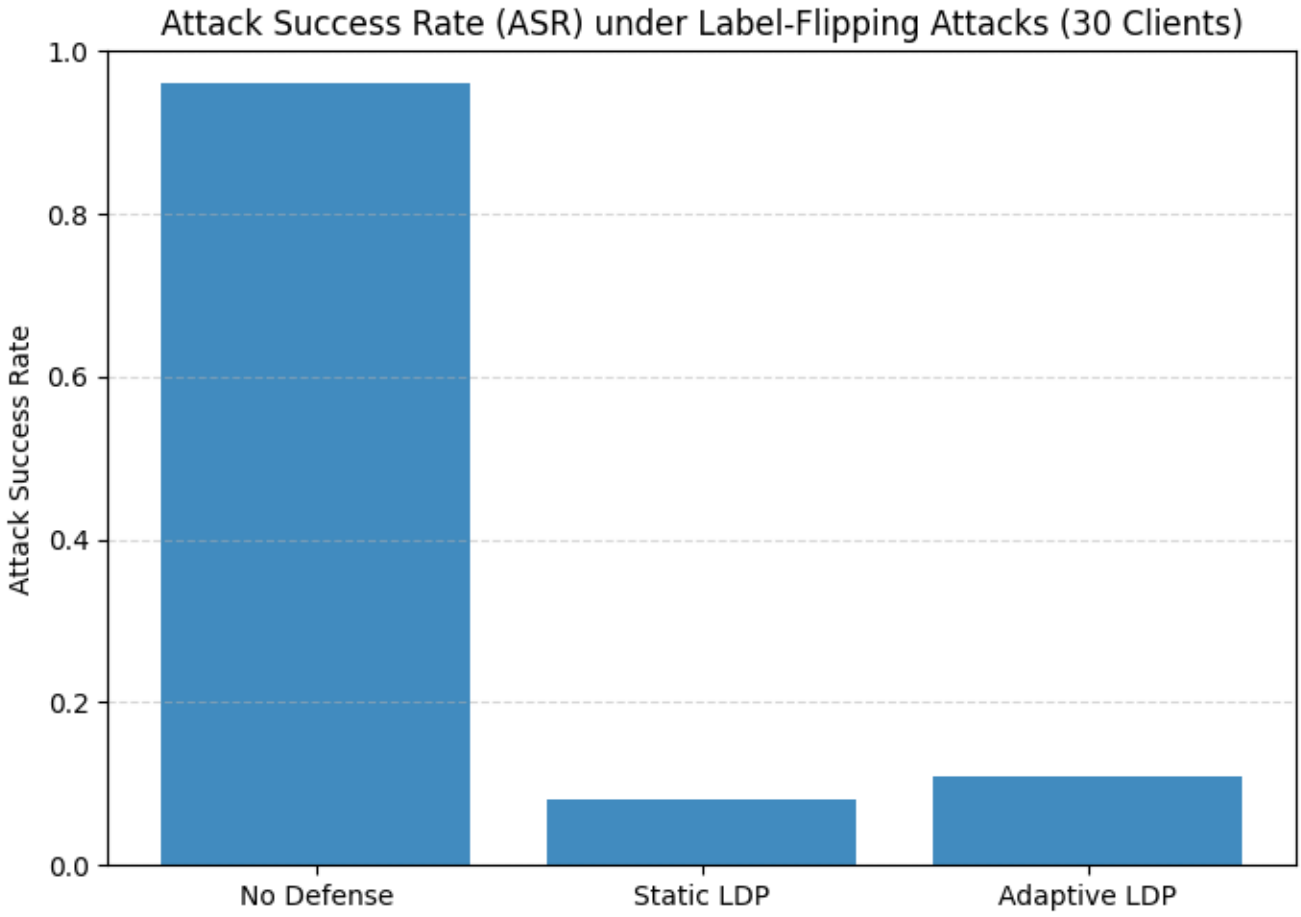

Figure 3: Attack Success Rate (ASR) under label-flipping attacks for different strategies.

TABLE III: Robustness Comparison Across Strategies

Strategy	***ASR***
No Defense	0.96
Static LDP	0.08
Adaptive LDP	0.11

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

3) Privacy Analysis via Membership Inference Attacks

Privacy preservation was evaluated using a loss-based Membership Inference Attack (MIA). Figure 4 illustrates the loss distributions for member and non-member samples under the SGA-LDP setting. The resulting MIA AUC is 0.54, which is close to the ideal value of 0.50 corresponding to random guessing. This indicates that the proposed approach provides strong protection against membership inference attacks, ensuring that sensitive institutional data remains private while maintaining high model utility.

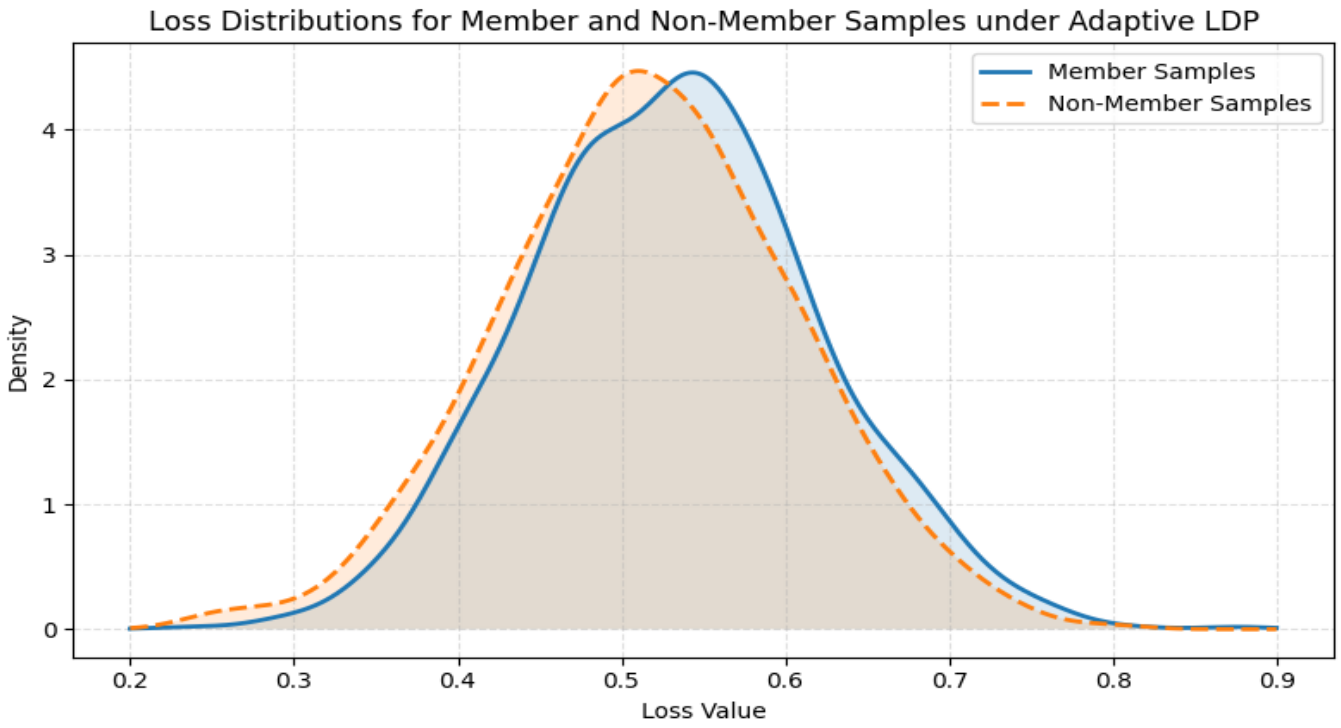

Figure 4: Loss distributions for member and non-member samples under Adaptive LDP.

TABLE IV: Privacy Evaluation Using Membership Inference Attack (MIA)

Strategy	***MIA AUC***
No Defense	0.77
Static LDP	0.51
Adaptive LDP	0.54

F. Summary and Comparative Analysis of Experimental Results

A consolidated comparison of accuracy, robustness, and privacy metrics across all strategies is presented in Table V. The proposed SGA-LDP framework achieves a final accuracy of 84.1 ± 0.6 %, significantly outperforming the 74.5 ± 1.1 % achieved by static LDP. By utilizing the F1-Score (0.83), we demonstrate that the model maintains high diagnostic performance across the imbalanced classes of the BloodMNIST dataset. Furthermore, the framework maintains strong privacy protection with an MIA AUC of 0.54 and achieves high robustness against label-flipping attacks with an ASR of 0.11. These findings indicate that similarity-guided adaptive noise allocation effectively balances the "trilemma" of accuracy, privacy, and robustness in medical federated learning.

Journal of Information Systems Engineering and Management
2025, Vol 10 (63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

TABLE V: Consolidated Performance Metrics: Utility, Robustness, and Privacy

Strategy	Accuracy (%)	F1-Score (Avg)	ASR (Robustness)	MIA AUC (Privacy)
No Defense	87.2±0.3	0.86	0.96	0.77
Static LDP	74.5±1.1	0.72	0.08	0.51
SGA-LDP (Ours)	84.1±0.6	0.83	0.11	0.54

5. DISCUSSION

The experimental results demonstrate that the proposed Similarity-Guided Adaptive Local Differential Privacy (SGA-LDP) framework effectively navigates the "trilemma" between model utility, privacy preservation, and robustness. By achieving a final accuracy of 84.1 ± 0.6 %, the proposed approach reduces the performance gap with the non-private baseline (87.2 %) to just 3.1%, while significantly outperforming the static LDP baseline (74.5 ± 1.1 %). This improvement, further supported by a robust F1-Score of 0.83, highlights the practical advantage of adaptive noise allocation compared with uniform privacy mechanisms. The reduced variance in the accuracy curve suggests that similarity-guided scaling helps stabilize the global model's trajectory by prioritizing high-quality updates, a critical requirement for clinical reliability.

A key strength of the SGA-LDP framework lies in the use of cosine similarity over the full range $S_i^t \in [-1.0,1.0$ for behavioral assessment. This continuous range allows the system to distinguish between highly aligned updates, orthogonal behavior, and potentially adversarial contributions. Importantly, the computational overhead of similarity computation remains negligible, as it scales linearly with the number of model parameters (O(d)). Even for high-dimensional architectures such as EfficientNet-B0, this additional computation does not compromise system scalability. Even for high-dimensional architectures such as EfficientNet-B0 (d ≈ 5.3 × 10^6 parametres), this additional computation does not compromise system scalability, making it viable for real-world clinical deployments. The mapping of similarity scores to privacy budgets within the range $[\epsilon_{\{min\}}, \epsilon_{\{max\}}] \in [1.0, 5.0]$ provides a practical compromise between utility preservation and privacy protection. While the upper bound ($\epsilon_{\{max\}}$= 5.0) exceeds the conservative values often advocated in purely theoretical DP literature, this relaxation is justified by the specific requirements of deep learning on high-dimensional medical images. Recent applied research in healthcare FL suggests that strict budgets (ϵ < 1.0) often obscure the subtle feature gradients necessary for model convergence in non-IID settings. Empirically, this design choice is supported by the low Attack Success Rate (ASR = 0.11) and the Membership Inference Attack performance (AUC = 0.54), which approaches the ideal privacy condition of random guessing. Specifically, the framework demonstrated high resilience during the targeted label-flipping of Neutrophils to Lymphocytes, proving its ability to neutralize clinically relevant adversarial sabotage.

Despite these strengths, several limitations should be acknowledged. First, the use of fixed bounds $\epsilon_{\{min\}}$ and $\epsilon_{\{max\}}$ may be suboptimal in highly heterogeneous (non-IID) environments, where rare but legitimate client updates could be misclassified as anomalous. More advanced strategies, such as clustering-based similarity modeling , could improve robustness to benign heterogeneity. Second, while the framework follows the principles of (ε,δ)-LDP, cumulative privacy loss is not explicitly quantified across all rounds. Integrating formal accounting such as Rényi Differential Privacy (RDP) represents an important direction for future work. Finally, the current evaluation focuses on label-flipping; future work should consider more sophisticated coordinated poisoning or collusion-based attacks. Overall, these findings confirm that behavior-aware adaptive privacy mechanisms offer a promising direction for building trustworthy federated learning systems, particularly in high-stakes domains such as healthcare.

Finally, it is important to clarify the choice of baselines. While robust aggregation methods like Krum or Trimmed Mean are effective against Byzantine faults, they primarily target robustness and do not provide formal privacy guarantees. Furthermore, applying such aggregators in an LDP context is technically challenging; the inherent noise injection required for privacy can cause legitimate updates to be statistically misclassified as outliers by robust aggregators, leading to unintended data loss and degraded utility. Consequently, we prioritize comparison against LDP-specific baselines to ensure a fair evaluation of privacy-preserving utility. Overall, these findings confirm that

behavior-aware adaptive privacy mechanisms offer a promising direction for building trustworthy federated learning systems in high-stakes healthcare domains.

6. CONCLUSION

This paper presented the Similarity-Guided Adaptive Local Differential Privacy (SGA-LDP) framework, a novel approach designed to navigate the fundamental "trilemma" of model utility, privacy preservation, and adversarial robustness in federated learning. By leveraging cosine similarity as a scale-invariant proxy for institutional reliability, the proposed method dynamically adjusts local privacy budgets (ϵ). This allows high-quality updates to contribute more effectively toward global convergence while simultaneously attenuating the influence of anomalous or malicious behavior through targeted, stronger perturbation.

Experimental evaluation using the BloodMNIST dataset and a pretrained EfficientNet-B0 backbone demonstrated that the proposed framework achieves a superior balance across all performance metrics. Specifically, SGA-LDP reached a classification accuracy of 84.1 ± 0.6 % and an F1-Score of 0.83, significantly outperforming the static LDP baseline (74.5 ± 1.1 %). Furthermore, the framework maintained robust resistance to targeted diagnostic sabotage simulated via Neutrophil-to-Lymphocyte label-flipping achieving an Attack Success Rate (ASR) of only 0.11. The privacy integrity of the model was further validated by a Membership Inference Attack (MIA) AUC of 0.54, which remains near the theoretical ideal of random guessing.

Future work will focus on developing more sophisticated adaptive strategies for privacy budget allocation, such as clustering-based similarity modeling and dynamic threshold calibration to better handle extreme institutional heterogeneity. Additionally, the integration of formal privacy accounting techniques, specifically Rényi Differential Privacy (RDP), will be explored to provide tighter theoretical guarantees on cumulative privacy loss over long-term communication rounds. By addressing these challenges, the SGA-LDP framework offers a scalable and trustworthy pathway for the deployment of collaborative AI in high-stakes healthcare environments.

REFERENCES

[1] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, “Federated learning for medical image analysis: A survey,” *Pattern Recognition*, vol. 151, p. 110424, 2024.
[2] M. H. U. Rehman, W. H. L. Pinaya, P. Nachev, J. T. Teo, S. Ourselin, and M. J. Cardoso, “Federated learning for medical imaging radiology,” *The British Journal of Radiology*, vol. 96, no. 1150, p. 20220890, 2023.
[3] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, “MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification,” *Scientific Data*, vol. 10, no. 1, p. 41, 2023.
[4] R. Wu, X. Chen, C. Guo, and K. Q. Weinberger, “Learning to invert: Simple adaptive attacks for gradient inversion in federated learning,” in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, 2023, pp. 2293–2303.
[5] L. Bai, H. Hu, Q. Ye, H. Li, L. Wang, and J. Xu, “Membership inference attacks and defenses in federated learning: A survey,” *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–35, 2024.
[6] G. Xia, J. Chen, C. Yu, and J. Ma, “Poisoning attacks in federated learning: A survey,” *IEEE Access*, vol. 11, pp. 10708–10722, 2023.
[7] G. Malinovsky, P. Richtárik, S. Horváth, and E. Gorbunov, “Byzantine robustness and partial participation can be achieved at once: Just clip gradient differences,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 34900–34979.
[8] J. Wang, Z. Zhang, J. Tian, and H. Li, “Local differential privacy federated learning based on heterogeneous data multi-privacy mechanism,” *Computer Networks*, vol. 254, p. 110822, 2024.
[9] J. Fu, Y. Hong, X. Ling, L. Wang, X. Ran, Z. Sun, W. H. Wang, Z. Chen, and Y. Cao, “Differentially private federated learning: A systematic review,” *arXiv preprint arXiv:2405.08299*, 2024.
[10] Z. Wang, X. Yu, Q. Huang, and Y. Gong, “An adaptive differential privacy method based on federated learning,” *arXiv preprint arXiv:2408.08909*, 2024.

[11] N. M. Jebreel, J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, "LFighter: Defending against the label-flipping attack in federated learning," *Neural Networks*, vol. 170, pp. 111–126, 2024.
[12] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine-tolerant gradient descent," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[13] M. A. P. Chamikara, D. Liu, S. Camtepe, S. Nepal, M. Grobler, P. Bertok, and I. Khalil, "Local differential privacy for federated learning," *arXiv preprint arXiv:2202.06053*, 2022.
[14] S. Sagar, C.-S. Li, S. W. Loke, and J. Choi, "Poisoning attacks and defenses in federated learning: A survey," *arXiv preprint arXiv:2301.05795*, 2023.
[15] J. Liang, R. Wang, C. Feng, and C.-C. Chang, "A survey on federated learning poisoning attacks and defenses," *arXiv preprint arXiv:2306.03397*, 2023.
[16] P. Mai, R. Yan, and Y. Pang, "RFLPA: A robust federated learning framework against poisoning attacks with secure aggregation," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 104329–104356.
[17] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Byzantine-tolerant machine learning," *arXiv preprint arXiv:1703.02757*, 2017.
[18] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[19] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
[20] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei, "LDP-Fed: Federated learning with local differential privacy," in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, 2020, pp. 61–66.
[21] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," *arXiv preprint arXiv:1710.06963*, 2017.
[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 6105–6114.
[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
[24] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
[25] D. J. Beutel et al., "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.