

FedRE: Robust and Effective Federated Learning with Privacy Preference

Tianzhe Xiao

School of Computer Science and Technology, Huazhong University of Science and Technology
Wuhan, China
d202381469@hust.edu.cn

Yichen Li

School of Computer Science and Technology, Huazhong University of Science and Technology
Wuhan, China
ycli0204@hust.edu.cn

Yu Zhou

Ant Group, Chongqing Ant Consumer Finance Co., Ltd
Chongqing, China
zy344525@myxiaojin.cn

Yining Qi*

School of Computer Science and Technology, Huazhong University of Science and Technology
Wuhan, China
qiyining@hust.edu.cn

Yi Liu

Ant Group, Chongqing Ant Consumer Finance Co., Ltd
Chongqing, China
larry.liuy@myxiaojin.cn

Wei Wang

Ant Group, Chongqing Ant Consumer Finance Co., Ltd
Chongqing, China
wangshi.ww@myxiaojin.cn

Haozhao Wang

School of Computer Science and Technology, Huazhong University of Science and Technology
Wuhan, China
hz_wang@hust.edu.cn

Yi Wang

Ant Group, Chongqing Ant Consumer Finance Co., Ltd
Chongqing, China
haonan.wy@myxiaojin.cn

Ruixuan Li*

School of Computer Science and Technology, Huazhong University of Science and Technology
Wuhan, China
rxli@hust.edu.cn

Abstract

Despite Federated Learning (FL) employing gradient aggregation at the server for distributed training to prevent the privacy leakage of raw data, private information can still be divulged through the analysis of uploaded gradients from clients. Substantial efforts have been made to integrate local differential privacy (LDP) into the system to achieve a strict privacy guarantee. However, existing methods fail to take practical issues into account by merely perturbing each sample with the same mechanism while each client may have their own privacy preferences on privacy-sensitive information (PSI), which is not uniformly distributed across the raw data. In such a case, excessive privacy protection from private-insensitive information can additionally introduce unnecessary noise, which may degrade the model performance. In this work, we study the PSI within data and develop *FedRE*, that can simultaneously achieve robustness and effectiveness benefits with LDP protection. More specifically, we first define PSI with regard to the privacy preferences of each client. Then, we optimize the LDP by allocating less privacy budget to gradients with higher PSI in a layer-wise manner, thus providing a stricter privacy guarantee for PSI. Furthermore, to

mitigate the performance degradation caused by LDP, we design a parameter aggregation mechanism based on the distribution of the perturbed information. We conducted experiments with text tamper detection on T-SROIE and DocTamper datasets, and FedRE achieves competitive performance compared to state-of-the-art methods.

CCS Concepts

• **Security and privacy** → *Distributed systems security*.

Keywords

Federated Learning, Local Differential Privacy, Privacy Preference, Privacy-Sensitive Information

ACM Reference Format:

Tianzhe Xiao, Yichen Li, Yu Zhou, Yining Qi, Yi Liu, Wei Wang, Haozhao Wang, Yi Wang, and Ruixuan Li. 2025. FedRE: Robust and Effective Federated Learning with Privacy Preference. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25)*, June 30–July 3, 2025, Chicago, IL, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3731715.3733336>

1 Introduction

Federated Learning (FL) has emerged as a basic paradigm that enables multiple parties to jointly train a model through the aggregation of parameters without sharing their private dataset [19, 23]. Due to the benefits of preserving privacy and communication efficiency, FL has been widely deployed in various applications, such as smart healthcare [1, 27] and finance analysis [4, 21, 38].

However, FL is not always impervious to security threats. A notable threat namely *gradient leakage attack*, aims to infer sensitive information from the shared model updates (gradients) [18]. Then,

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR '25, Chicago, IL, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1877-9/2025/06

<https://doi.org/10.1145/3731715.3733336>



Figure 1: This illustration depicts the varied privacy preferences of different clients. Client A pays close attention to iconographic elements within the data, such as stamps, bar codes, and QR codes, and highlights these privacy-sensitive regions with a yellow border. In contrast, Client B is more concerned with textual content, including numerical values and phone numbers, marking these sections with a blue frame.

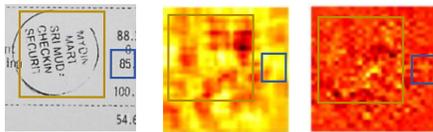


Figure 2: Different derivatives of different layer networks on input images (left : original image, middle and right : derivatives on different layers).

the malicious participant can exploit this information to reconstruct the private data or infer properties about it, leading to the violation of local privacy [8, 48].

To address this issue, several defense methods have been proposed to mitigate the risk of gradient leakage attacks in FL [10, 32]. LDP-Fed [30] is proposed to optimize LDP for the FL system which ensures a lightweight and quantifiable measure for privacy preservation. The authors in [5] aim to solve performance degradation in FL with user-level DP and employ regularization and sparsification techniques to local updates. FedDPA is proposed in [40] to study the differential privacy in the personalized FL with dynamic fisher personalization and adaptive constraint. PrivateRec [20] focuses on federated recommendation scenarios and is devoted to achieving better utility in online serving under a DP guarantee.

While these approaches have achieved great success, they are generally designed for the scenario where the same privacy protection mechanism is applied to all samples, overlooking two critical realities: First, clients are likely to have distinct privacy preferences, meaning varied privacy-sensitive information (PSI) for protection. The complexity of this issue arises from the attributes of privacy preference, which are not always explicitly observed in the data. PSI of privacy preferences may not be as overt as specific pixel regions within an image or particular words within a text. Rather, they may hinge on the overarching structure of the data, encompassing semantic information. Second, PSI is not uniformly distributed across

data. Implementing a sample and indiscriminate privacy protection approach across all data samples can lead to the introduction of unnecessary noise, which can adversely affect model performance.

Considering that we are the first to propose and explore the protection of privacy preferences of clients in FL, in this paper, we seek to explore a foundational privacy protection scenario, namely *privacy-sensitive regions* in images. We illustrate this concept in Fig. 1. To explore this scenario, we have devised a robust and effective strategy for the preservation of PSI in regions. Local Differential Privacy [2, 6] have demonstrated great success against gradient leakage attacks by perturbing samples with the privacy budget. A direct idea to solve the problem is to apply these LDP-based methods to perturb the designated privacy-sensitive regions with a lower privacy budget, which provides a stricter privacy guarantee. Despite the simplicity of such an approach, applying direct pixel-level perturbation results in the loss of critical feature information, which in turn compromises model performance [46]. This presents a challenge in striking the optimal balance between the robustness of the privacy protection and the effectiveness of the model accuracy.

To tackle this challenge, we propose FedRE - which can ensure both robustness and effectiveness benefits in the FL system. Figure 2 shows the derivatives calculated by different network layers on the same input image. We observe that if the sensitive regions selected by the client are different (yellow or blue boxes), the derivative values accumulated by different regions in different layers of the network are different. Inspired by [32], we study the layer-wise information leakage from the gradients, using the sensitivity of gradient changes regarding the PSI region to quantify the leakage risk. Then, we allocate different privacy budgets to perturb the gradients of each layer guided by the sensitivity. To mitigate the adverse effects that local gradient perturbation may have on the performance of the global model, we introduce a new aggregation mechanism. Upon receiving gradients from local clients, the server employs a publicly available dataset to evaluate the sensitivity of these gradients. The global model will favor aggregating less sensitive local gradients, which can reduce the infusion of noise from the local perturbed gradients, thereby preserving the effectiveness of the global model.

To verify the effectiveness of our method, we first manually annotate the PSI region of two real-world datasets: T-SROIE and DocTamper. Based on these datasets, extensive experiments have been done and show that the proposed FedRE enables more accurate and robust models relative to state-of-the-art baselines. The major contributions of this paper are summarized as follows:

- We propose and formally define the concept of privacy preferences in the context of federated learning, highlighting the need to protect privacy-sensitive information (PSI) in privacy-sensitive regions. Our definition accounts for the diverse and unique privacy concerns of different clients, acknowledging that PSI can vary significantly between data and clients.
- We introduce **FedRE**, a novel method that integrates local differential privacy (LDP) in a layer-wise manner to provide tailored privacy protection for PSI. Our approach judiciously allocates the privacy budget across layers based on the sensitivity of the gradients to PSI, allowing for a more nuanced

and effective privacy guarantee without substantially compromising on the model’s performance.

- We conduct extensive experiments on the annotated PSI regions of the T-SROIE and DocTamper datasets to validate the effectiveness of our proposed method. Our empirical results demonstrate that FedRE achieves superior performance in terms of robustness and accuracy compared to existing state-of-the-art methods, thereby confirming the practical utility of our approach in real-world FL scenarios.

2 Related Work

Federated Learning. Federated Learning is a distributed machine learning approach that enables multiple entities to collaboratively train a model without directly sharing raw data, thereby preserving data privacy and security [12, 16]. Key research directions in this field encompass algorithm optimization for efficient learning [31, 36], security and privacy enhancements [26, 41], system and architectural design for scalability [25], graph learning [9, 24], continual learning [17, 22], and incentive mechanisms to encourage and select participation [29, 37].

Large corporations have shown significant interest in federated learning for various applications. For instance, in the financial sector, banks can leverage federated learning for enhanced fraud detection [39, 47], allowing them to share insights from their models without revealing sensitive transaction data. Other applications include personalized recommendations in e-commerce and news [20, 41], medical research in healthcare [27, 28], and device optimization in manufacturing [7, 14]. This approach enables the collective enhancement of various capabilities while maintaining strict data privacy.

Gradient leakage attack. Federated learning is susceptible to malicious attacks, including gradient leakage [35], model inversion [15], and membership inference [44] attacks. Gradient leakage is particularly harmful as it can reveal extensive information from the victim’s training data. This attack method involves initializing pseudo training data and labels, and optimizing them to mirror real gradients. As the pseudo and real gradients converge, the pseudo data begins to reflect the properties of the actual private data.

The method proposed in [48] can effectively attack not only computer vision tasks but also natural language processing tasks. Subsequent work has improved in areas such as initialization with prior knowledge [11], ground-truth label extraction [45], faster optimizer [8], and regularization terms [8, 42]. These improvements enable more effective gradient leakage attacks on larger batch sizes, higher resolutions, and more complex models (such as ViT) [42]. Therefore, understanding and mitigating gradient leakage attacks is crucial due to their potential to cause significant harm.

Privacy Protection. Privacy protection in machine learning encompasses various techniques aimed at safeguarding sensitive information. Differential privacy is a technique that introduces noise to the gradients before they are shared, thereby limiting the amount of information that can be inferred from them [30, 34]. This method provides a mathematical guarantee of privacy but at the cost of model accuracy. Secure aggregation is another technique where the gradients are encrypted in a way that allows the server to compute their sum without being able to decrypt individual gradients [3].

This method provides robust security guarantees but requires more computational resources. Homomorphic encryption is a cryptographic technique that allows computations to be performed on encrypted data without decrypting it, providing another layer of security [43].

However, a uniform privacy protection mechanism based on these techniques is deployed across all samples, ignoring the PSI distribution within data and the privacy preference.

3 Methodology

We first formulate privacy preference scenarios and propose the robust and effective FedRE. Then, we present a scalable algorithm and provide rigorous analytical results to show the efficiency of the proposed method.

3.1 Problem Formulation

FL Procedures. Our work is developed based on the paradigm of Federated Averaging (FedAvg) algorithm. FedAvg, introduced by Google in 2016 [23], is a seminal work in the domain of federated learning. Initially designed for privacy-preserving machine learning on mobile devices, FedAvg transcends mobile applications to enable collaborative model training across multiple institutions. In such federated settings, institutions maintain the privacy of their local data while collectively training a global model through the exchange of model updates, not raw data.

Based on FedAvg, we aim to collaboratively train a global model for K total clients in FL. We consider each client k can only access to his local private dataset $D_k := \{x_i, y_i\}$, where x_i is the i -th input data sample and $y_i \in \{1, 2, \dots, C\}$ is the corresponding label of x_i with C classes. The global dataset is considered as the composition of all local datasets $D = \sum_{k=1}^K D_k$. The objective of the FL learning system is to learn a global model w that minimizes the total empirical loss over the entire dataset D :

$$\min_w \mathcal{L}(w) := \sum_{k=1}^K \frac{|D_k|}{|D|} \mathcal{L}_k(w),$$

$$\text{where } \mathcal{L}_k(w) = \frac{1}{|D_k|} \sum_{i=1}^{|D_k|} \mathcal{L}_{CE}(w; x_i, y_i), \quad (1)$$

where $\mathcal{L}_k(w)$ is the local loss in the k -th client and \mathcal{L}_{CE} is the cross-entropy loss function that measures the difference between the prediction and the ground truth labels.

Local Differential privacy. In the context of FL, clients collaborate to train a global model under the constraint that each client’s data remains local. While this protects the raw data, the gradients shared during training can still leak sensitive information. Traditional DP requires a central trusted party which is often not realistic. To remove that limitation, local differential privacy (LDP) has been proposed. The definition of (ϵ, δ) -LDP is given as below:

Definition 3.1. A perturbation algorithm M satisfies (ϵ, δ) -Local Differential Privacy $((\epsilon, \delta)$ -LDP) if, for any pair of adjacent datasets D and D' , and for all possible output subsets S , the following inequality holds:

$$\Pr[M(D) \in S] \leq e^\epsilon \Pr[M(D') \in S] + \delta \quad (2)$$

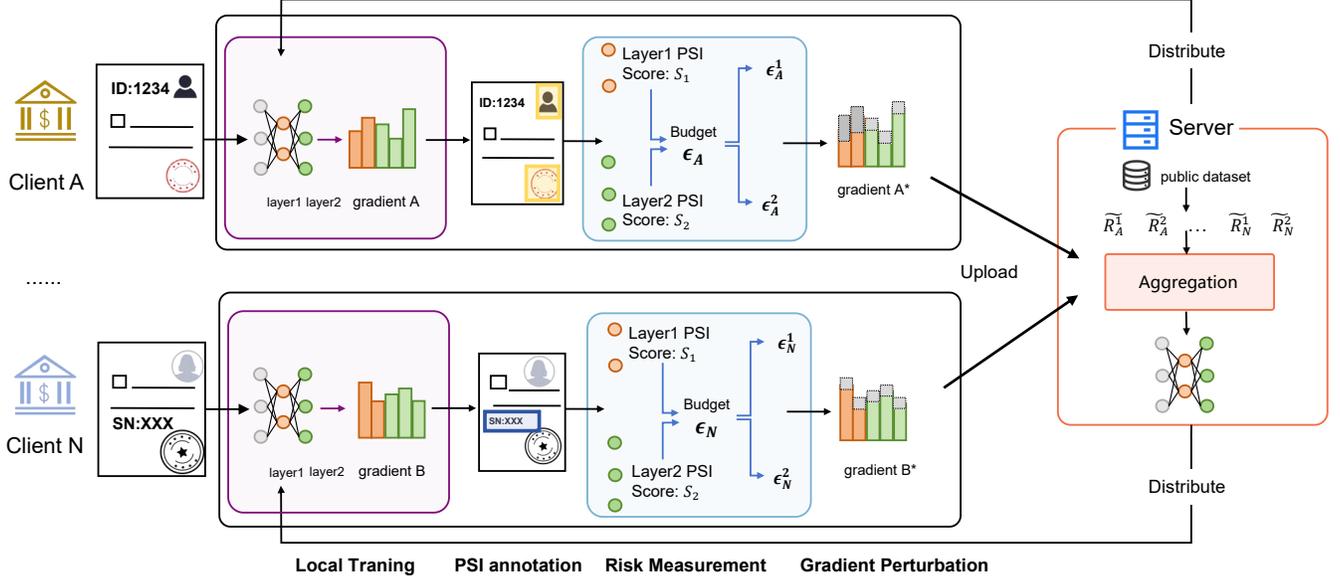


Figure 3: The overall architecture of FedRE. When clients want to train a model cooperatively, each of them first trains locally on the last round’s global model to get the original gradient. Then based on annotated PSI, we can measure the risk of PSI leakage at each layer by calculating their PSI scores, and the privacy budget ϵ will be rationally allocated to each layer accordingly. The original gradient will then be perturbed according to the privacy budget of each layer correspondingly and uploaded to the server for aggregation. Finally, in order to reduce the degradation caused by perturbation, the server aggregates all the gradients based on the possible distribution of the perturbation information known using the public dataset and distributes the updated model to all clients for the training of the next round.

where ϵ is the privacy budget of M , which quantifies the privacy protection level, and δ is the probability of the privacy guarantee being violated. A smaller value for ϵ indicates a smaller gap between two probabilities and thus a stronger privacy.

Threat Models. We assume that in the entire federated learning environment, local clients only upload their trained gradients. The central server is intrusted and may initiate a gradient inversion attack while aggregating gradients. This attack maliciously infers the client’s private training data by comparing the model broadcast in the previous round with the gradient uploaded by the client in the current round, attempting to restore as much detailed information in D_i as possible. If an eavesdropper in the communication channel intercepts the interaction information between the client and the central server, it also can launch the same attack.

3.2 FedRE: Protection of Privacy Preference

The key idea of FedRE is to employ a layer-wise local differential privacy mechanism tailored to local privacy preferences, ensuring precise protection of Privacy-Sensitive Information. More specifically, we first compute the sensitivity of the PSI in the privacy-sensitive region. Then we allocate privacy budgets based on the sensitivity of gradients at different network layers to PSI, reducing the noise from the perturbation of privacy-insensitive information. Moreover, with a novel aggregation mechanism on the server side, FedRE gives priority to the gradients with less sensitive information, minimizing the impact of perturbation on global model performance. The

workflow of the proposed framework is shown in Algorithm 1 and Fig. 3 illustrates the FedRE approach.

3.2.1 Measure of Sensitive Private Information. To quantify the sensitivity of private information of privacy preference, we can re-frame the privacy leakage as an issue of the model gradient’s sensitivity to input data. When the model calculates gradients, some parameters may be particularly responsive to changes in the private-sensitive regions. The variability in these parameters could be greater, suggesting that they hold more information from those regions, which could increase the risk of privacy breaches during gradient inversion attacks. Inspired by this insight, we employ the *Jacobian* matrix of the gradient concerning the input as a tool to gauge the sensitivity of different gradient segments to the input data:

$$J_l(x) = \frac{\partial g_l(x)}{\partial x} = \frac{\partial}{\partial x} \left[\frac{\partial l(x, y; w)}{\partial w_l} \right] \quad (3)$$

where $g_l(\cdot)$ is equivalent to the partial derivative of the loss function $l(\cdot)$ with respect to the parameters w in the l -th layer.

Then, we extract the privacy-sensitive regions from the data, and for each pixel within the region, we align the values across different pixel channels with the *frobenius-norm* since Jacobians are compared across layers with different sizes and *frobenius-norm* will consider all dimensions of the data. Assuming the dimensions of the privacy-sensitive region are $(w \times h \times c)$, where c is the number of channels and $w \times h$ represents the region size, we can calculate

Algorithm 1: FedRE

Input: T : communication round; K : client number; η : learning rate; $\{D_t\}_{t=1}^K$: distributed dataset with K clients; w : parameter of the model; ϵ_l : privacy budget for l -th layer in model; $\sum\{\epsilon_t\}_{t=1}^l$: total privacy budget; \mathbb{D} : the public dataset.

- 1 Initialize the parameter w ;
- 2 **for** $t = 1$ **to** T **do**
- 3 Server randomly selects device subset S_t and send w
- 4 **for each selected client** $k \in S_t$ **in parallel do**
- 5 **for each layer** l **in the local model do**

Measure of Sensitive Private Information
 Compute the Jacobian matrix $J_l(x)$ of the gradient as the sensitivity with (3);
 Align the sensitivity of privacy-sensitive region $J_l^R(x)$ from different channels with (4);
 Compute the averaged PSI score S_l for each sample with (5);
- 6 **end**
- 7

Local Differential Privacy with PSI Score
 Clip each gradient g to g^c with (7); Perturb each gradient g^c with (8) and (9).
- 8 **end**
- 9 Send the model w^k back to the server.
- 10 **end**
- 11 **At server side**

Parameter Aggregation Mechanism
 Normalize the weight $\hat{\alpha}$ for the aggregation with the public dataset with (10); Aggregate the local gradients with the weight to obtain the global model w with (11).
- 12 **end**
- 13 **end**

the aligned sensitivity of privacy-sensitive region $J_l^R(x)$:

$$J_l^R(x) = \begin{bmatrix} \|J_l(x)_{[a,b,:]} \|_F & \cdots & \|J_l(x)_{[a,b+h,:]} \|_F \\ \vdots & \ddots & \vdots \\ \|J_l(x)_{[a+w,b,:]} \|_F & \cdots & \|J_l(x)_{[a+w,b+h,:]} \|_F \end{bmatrix} \quad (4)$$

Given the client k , we compute the average of the aligned sensitivity of the privacy-sensitive region as the PSI score in the l -th layer of gradients:

$$S_l = \frac{1}{w * h} \sum_{i=1}^w \sum_{j=1}^h J_l^R(x)_{[i,j]} \quad (5)$$

3.2.2 Local Differential Privacy with PSI Score. Acquiring both the gradient g and PSI score S , the client starts figuring out the right amount of noise for differential privacy in a layer-wise manner. While a smaller privacy budget represents a stricter protection mechanism, we allocate the privacy budget for each layer according to the PSI score S_l , thereby ensuring a balance between the

effectiveness of model performance and privacy protection.

$$\epsilon_l = \frac{\epsilon \times \frac{1}{S_l}}{\sum_{t=1}^L \frac{1}{S_t}} \quad (6)$$

Where ϵ is the total privacy budget set by the client. To implement a differential privacy perturbation mechanism that complies with the privacy budget ϵ for gradients, we adopt the clipping and noise addition to ensure that the global model update is indistinguishable whether a particular sample is included in the learning process.

$$g_l^c = \min \left(1, \frac{C_l}{\|g_l\|} \right) \times g_l \quad (7)$$

Where C_l is the clipping threshold of l -th layer that controls the maximum contribution of a training sample to global update, g_l^c is the gradient after clipping. Besides, to make it potential attackers to infer specific information of any sample, noise adding is performed to satisfy the randomness requirement of DP. We take the Gaussian mechanism for gradient noise adding to ensure LDP. It adopts L_2 norm sensitivity, and adds zero-mean noise with variance $C_l^2 \sigma_l^2 \mathbf{I}$:

$$\mathcal{M}(g) = g_l^c + \mathcal{N}(0, C_l^2 \sigma_l^2 \mathbf{I}) \quad (8)$$

Where \mathbf{I} is an identity matrix and has the same size with g_l^c . σ_l is a noise multiplier computed by a privacy accountant and composition mechanism [33] for privacy budget ϵ_l , failure probability δ_l and communication rounds T .

$$\sigma_l = \frac{\sqrt{2T \ln(1/\delta_l)}}{\epsilon_l} \quad (9)$$

Theorem 1 (Simple Composition) Here we introduce Theorem 1 proposed by [13]. If \mathcal{M}_i is an (ϵ_i, δ_i) -differentially private (DP) mechanism, then the composition $(\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k)$ satisfies $(\sum_{i=1}^k \epsilon_i, \sum_{i=1}^k \delta_i)$ -DP.

Corollary 1 (DP Composition in FedRE) Denote the gradient of the network w with l -layers $G = [g_1, g_2, \dots, g_l]$, privacy budget for each layer $\epsilon = [\epsilon_1, \epsilon_2, \dots, \epsilon_l]$, $\sum_{i=1}^l \epsilon_i = \epsilon$, and probability of being violated for each layer $\delta = [\delta_1, \delta_2, \dots, \delta_l]$, $\sum_{i=1}^l \delta_i = \delta$. Given any gradient of the l -th layer, the proposed mechanism \mathcal{M}_l in (8) satisfies (ϵ_l, δ_l) -LDP. Then, the gradient of the w satisfies (ϵ, δ) -LDP.

3.2.3 Parameter Aggregation Mechanism. Considering that the local perturbed gradient will introduce noise to the aggregated global model, which may degrade the model performance, we develop a new Perturbation Distribution Aware Parameter Aggregation Mechanism (**PDA-PAM**) that can be aware of the distribution of client's parameter perturbation, enabling the server to aggregate the clean gradients (with less perturbation during the model training) from local clients. Assuming that the server has access to the category of local data, then the server employs a public dataset and computes the PSI score of each layer (defined in 3.2.1) of each local model w^k with the public dataset \mathbb{D} and then normalizes it into the weight:

$$\hat{\alpha}_l^t = \frac{e^{S_l(w_l^t; \mathbb{D})}}{\sum_{t=1}^K e^{S_l(w_l^t; \mathbb{D})}} \quad (10)$$

which guarantees that $\sum_{t=1}^K \hat{\alpha}^t = 1$. Finally, the server aggregates the local gradients with the PSI score on the public dataset to obtain the global model w for the next communication round:

$$w_l = \tilde{w}_l - \frac{\eta}{K} \sum_{t=1}^K \frac{g_l^t}{\hat{\alpha}_l^t} \quad (11)$$

where \tilde{w}_l denotes the l -th layer of the global model in the last communication round. Here the server prefers to aggregate the gradients with less sensitivity thus the global model can gain more effective information.

3.3 Experimental Results

Training performance. We analyze the utility of different methods on two datasets. As shown in Table 1, the complexity and diversity of the DocTammer dataset could pose additional challenges for maintaining high accuracy, especially when the privacy constraint is strict ($\epsilon = 10$). Nonetheless, FedRE consistently performs on par or outperforms other methods across all metrics and datasets, indicating its robustness and adaptability.

Comparing the results across different ϵ values, we observe that an increase in the privacy budget (i.e., larger ϵ) leads to improved performance for all methods. This is expected, as a larger ϵ allows for less noise to be injected during the federated learning process, thus facilitating better model convergence.

The performance gains for FedRE are more notable on the more complex DocTammer dataset when the privacy budget is tight ($\epsilon = 10$). This underscores the effectiveness of FedRE’s PDA-PAM aggregation strategies, which are able to effectively handle the diverse and potentially conflicting perturbations present in the client models’ updates. By dynamically adjusting the aggregation based on the specific perturbations, FedRE is able to extract useful information even from heavily distorted updates, demonstrating its resilience in challenging conditions.

Defense performance. After we prove that FedRE can provide similar or even superior learning results compared to state-of-the-art DP-based FL mechanisms, we study if the sensitivity computation improves defense ability. In our experimental evaluation, FedRE’s sensitivity-driven privacy budget allocation strategy has demonstrated remarkable effectiveness in enhancing the defense capabilities against adversarial attacks in real-world privacy-preserving applications. Specifically, by identifying and prioritizing sensitive personally identifiable information within the data, FedRE is able to allocate more privacy budget to these critical regions. Fig. 4 shows an example of a real-life privacy-preserving application of FedRE, e.g., for the same privacy budget, by labeling the last three digits of the social security number as the PSI that need to be protected, FedRE can allocate more privacy budget to the areas that need it, thus successfully blurring the PSI recovered from the attack, and decreasing the likelihood of compromising the privacy information.

Quantitative results presented in Table 2 consistently show that FedRE outperforms other state-of-the-art DP-based FL mechanisms across various metrics and datasets. This superior performance can be attributed to FedRE’s budget allocation strategy, which focuses on protecting sensitive areas more rigorously. Consequently, the similarity between the recovered PSI and the original information is substantially reduced.



Figure 4: Images recovered from gradient after gradient leakage attack without FedRE and with FedRE under the same privacy budget. Assuming that the last three characters of an image containing a social security number are PSI, the left image is the original image, the center image is attacked without FedRE, and the right image is the effect of privacy budget reallocation using FedRE.

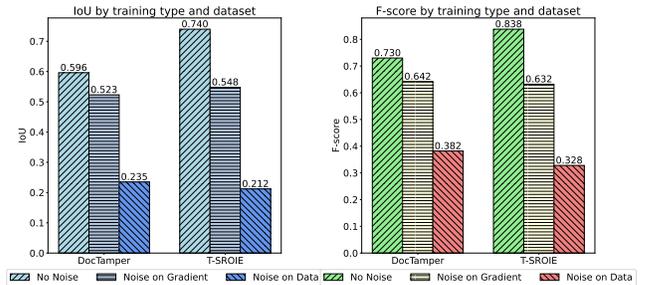


Figure 5: Comparison of the effects of no noise, adding noise to the gradient and adding noise to the raw data on the DocTammer and T-Sroie datasets for training, with the iou metric on the left and the f-score metric on the right.

The advantage of FedRE is even more pronounced when the privacy budget (ϵ) is set to a lower value, such as 10. This is because a stricter privacy budget encourages FedRE to allocate more noise to gradients that contain information related to PSI, thereby strengthening the defense.

The disparity in defense metrics between the T-Sroie and DocTammer datasets highlights FedRE’s adaptability to diverse data types. Given that all images in the T-Sroie dataset are grayscale invoices, while DocTammer contains more diverse and colorful images, the defense metrics on the DocTammer dataset exhibit a more pronounced numerical advantage. Nevertheless, FedRE maintains its superior performance, indicating its robustness across different data landscapes.

The Necessity of Gradient Perturbation for PSI Protection. To underscore the advantages of perturbing gradients for safeguarding private and sensitive information (PSI), as opposed to directly perturbing raw data at the pixel level, we delve deeper into the utility trade-offs between these two methods. Our initial approach involves directly infusing Gaussian noise into the privacy-sensitive regions of each image, while the alternative approach strategically introduces noise to the gradients during the training process. For a fair comparison, we meticulously adjust the noise intensities to ensure that after 2000 iterations of DLG attack, both methods exhibit comparable defense capabilities, as measured by similarity metrics.

As evident in Fig 5, while gradient perturbation introduces a modest performance decrement, pixel-level perturbation to raw data leads to a drastic deterioration in utility. This disparity stems from the nuanced requirements of tasks such as tamper detection,

Table 1: Comparison of Training Performance on Different Datasets.

ϵ	METHODS	T-SROIE				DocTAMPER			
		IoU	PRECISION	RECALL	F-SCORE	IoU	PRECISION	RECALL	F-SCORE
∞	CENTRAL	0.721±0.018	0.771±0.008	0.917±0.020	0.838±0.014	0.575±0.012	0.778±0.006	0.688±0.012	0.729±0.010
	LOCALSET	0.408±0.017	0.607±0.018	0.564±0.016	0.585±0.015	0.345±0.015	0.677±0.019	0.599±0.014	0.635±0.016
50	LDP-FED	0.523±0.019	0.654±0.017	0.681±0.018	0.667±0.018	0.498±0.018	0.716±0.018	0.652±0.017	0.682±0.019
	BLUR+LUS	0.577±0.020	0.689±0.019	0.639±0.019	0.663±0.019	0.514±0.019	0.707±0.020	0.626±0.018	0.664±0.020
	FEDRE	0.601±0.018	0.697±0.016	0.651±0.017	0.673±0.017	0.524±0.017	0.727±0.017	0.643±0.016	0.683±0.018
10	LDP-FED	0.430±0.016	0.592±0.021	0.594±0.015	0.593±0.017	0.424±0.014	0.647±0.022	0.574±0.013	0.608±0.018
	BLUR+LUS	0.439±0.017	0.611±0.023	0.606±0.016	0.609±0.018	0.432±0.015	0.657±0.024	0.583±0.014	0.618±0.019
	FEDRE	0.448±0.015	0.630±0.020	0.618±0.014	0.624±0.016	0.457±0.013	0.686±0.021	0.609±0.012	0.645±0.017

Table 2: Comparison of Defense Performance on Different Datasets.

ϵ	METHODS	T-SROIE				DocTAMPER			
		MSE \uparrow	SSIM \downarrow	PSNR \downarrow	LPIPS \uparrow	MSE \uparrow	SSIM \downarrow	PSNR \downarrow	LPIPS \uparrow
50	LDP-FED	0.022 ± 0.003	0.937 ± 0.005	48.976 ± 0.611	0.279 ± 0.014	0.027 ± 0.004	0.897 ± 0.008	46.375 ± 0.693	0.288 ± 0.017
	BLUR+LUS	0.020 ± 0.002	0.894 ± 0.005	49.978 ± 0.520	0.269 ± 0.013	0.025 ± 0.003	0.904 ± 0.007	47.877 ± 0.587	0.282 ± 0.016
	FEDRE	0.025 ± 0.004	0.808 ± 0.008	47.463 ± 0.727	0.294 ± 0.018	0.030 ± 0.004	0.888 ± 0.008	45.462 ± 0.813	0.303 ± 0.019
10	LDP-FED	0.327 ± 0.016	0.792 ± 0.010	45.985 ± 1.211	0.391 ± 0.022	0.422 ± 0.019	0.768 ± 0.011	42.986 ± 1.324	0.303 ± 0.023
	BLUR+LUS	0.375 ± 0.017	0.739 ± 0.009	45.987 ± 1.120	0.383 ± 0.021	0.390 ± 0.011	0.776 ± 0.010	43.984 ± 1.235	0.296 ± 0.022
	FEDRE	0.383 ± 0.014	0.676 ± 0.008	43.772 ± 1.313	0.412 ± 0.024	0.438 ± 0.020	0.653 ± 0.014	40.367 ± 1.421	0.324 ± 0.023

which heavily rely on intricate noise patterns and accurate color perception in the raw data. Notably, privacy-sensitive regions often coincide with areas that have undergone tampering, thus, applying noise to these overlapping zones disrupts the model’s ability to extract meaningful and effective knowledge from them. Consequently, the model’s capacity to accurately detect and classify tampering instances is significantly hindered.

The result also shows a dataset-specific trend. The T-SROIE dataset, being relatively smaller in size, appears to be more susceptible to the detrimental effects of noise-augmented training. Specifically, the introduction of noise during training leads to a far more pronounced reduction in IoU and F-score compared to the DocTamper dataset. This observation underscores the importance of tailoring privacy-preserving techniques to the unique characteristics of individual datasets, particularly their size and complexity, to ensure a balanced approach that safeguards privacy without compromising too much on utility.

The gradient perturbation approach offers a more flexible and targeted means of defense. *By perturbing gradients rather than the raw data, we can maintain a higher level of fidelity in the input images, allowing the model to better capture relevant features for downstream tasks.* This targeted intervention not only reduces the overall performance impact but also ensures that the model’s ability to detect tampering remains robust, even in the presence of privacy-preserving measures.

Parameter Aggregation Mechanism Gain. Fig. 6 presents a comparison between IoU results when parameters are aggregated

with and without the implementation of PDA-PAM in FedRE. The figure elucidates the impact of varying the privacy budget ϵ on the performance of the system, particularly under conditions where this budget is constrained.

As the privacy budget ϵ diminishes, the aggregation of parameters utilizing PDA-PAM from a larger number of clients is observed to compensate more effectively for the noise introduced by the differential privacy constraints. This phenomenon can be attributed to the diversity of Privacy-Sensitive Information (PSI) across different clients. Since the privacy budget ϵ_l for the same layer may vary among clients, those with a larger ϵ_l can offer better compensation for clients with a smaller ϵ_l , especially when the collective client count is high. When ϵ_l is lower, the compensation effect will be more obvious. However, due to the limitation of the fixed size of the data set we use for the experiment, this trend may become less pronounced once the number of clients reaches a certain level. This may be because the amount of information that may be provided by each additional client gradually decreases, which can be mitigated if the data set grows with the number of clients in real-world scenarios.

When the privacy budget is less stringent and the system comprises only a single client, the IoU values surpass those of centralized training without applying differential privacy. This outcome may be accredited to the clipping operation under a lower noise regime, which could potentially enhance gradient regularization. This observation implies that in real life, the addition of differential privacy does not always result in a worse performance.

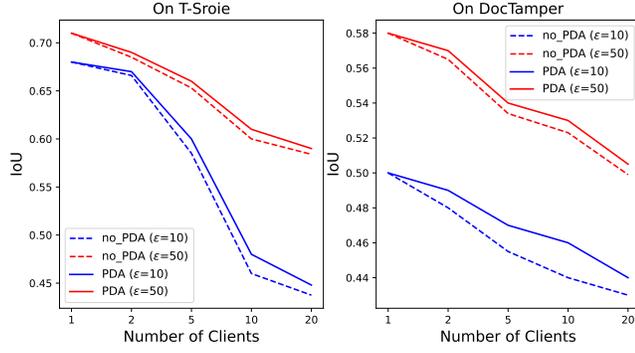


Figure 6: Iou gain of aggregation using PDA-PAM under the different number of clients and privacy overhead settings.

Clipping threshold. Table 3 and 4 are experiment results with respect to thresholds C_l in (7) on DocTamper and T-SROIE dataset. All experiments are conducted under the setting of $\epsilon = 50$, $lr = 0.005$, and clients number = 10. We can see that a suitable large threshold will not affect the training performance too much, and may even slightly improve performance, acting as a way of regularization. However, too small a threshold will lead to too little information contained in the gradient, resulting in the model being unable to learn during training.

Table 3: DocTamper Dataset Results for Threshold C_l .

C_l	IOU	PRECISION	RECALL	F-SCORE
0.20	0.535±0.021	0.701±0.016	0.707±0.017	0.702±0.022
0.15	0.524±0.017	0.727±0.017	0.643±0.016	0.683±0.018
0.07	0.283±0.020	0.534±0.016	0.344±0.017	0.428±0.023
0.05	0.260±0.013	0.313±0.014	0.501±0.015	0.384±0.017
0.03	0.000	0.000	0.000	0.000

Table 4: T-SROIE Dataset Results for Threshold C_l .

C_l	IOU	PRECISION	RECALL	F-SCORE
0.25	0.593±0.020	0.687±0.020	0.649±0.018	0.671±0.021
0.20	0.601±0.018	0.697±0.016	0.651±0.017	0.673±0.017
0.15	0.322±0.017	0.378±0.018	0.623±0.015	0.465±0.015
0.14	0.311±0.018	0.336±0.018	0.601±0.016	0.435±0.014
0.13	0.000	0.000	0.000	0.000

Computational Overhead Analysis. The process of calculating the PSI score involves three processes, corresponding to (3), (4), and (5) in the paper.

Equation (3) corresponds to the process of evaluating the Jacobian matrix, which involves two sub steps; the first step is to obtain the gradient by calculate derivative of the loss function with respect to the model weights, and the second step is to obtain the Jacobian matrix by taking the derivative of gradient with respect to the input data of the model. The time complexity of the first step is $O(F)$, F represents the total number of floating-point computations performed by the model, and the space complexity is $O(N)$, N represents the number of parameters of the model. This step is the same as the original gradient computation process in model training task, and thus can utilize the intermediate results

generated during training without incurring any additional time and space overheads. The time complexity of the second step is $O(FD)$, where D is the number of features in the input data and the space complexity is $O(DN)$. Equation (4) corresponds to aligning the sensitivity of the privacy-sensitive region, assuming that the total number of pixel points in the privacy-sensitive region is p , the time complexity is $O(p)$ and the space complexity is $O(p)$. Equation (5) corresponds to the calculation of the average PSI score within the privacy-sensitive region, with a time complexity of $O(p)$ and a space complexity of $O(p)$.

p is generally much smaller than FD and DN , so in summary the time complexity of the algorithm is $O(FD)$ and the space complexity is $O(DN)$.

While the calculation of PSI scores has some overhead, in practice it is not necessary to calculate PSI scores for every training data, but only when dealing with data with different content formats. Financial data generally have several fixed formats, such as contract, invoice, normal page, receipt, etc.. There are large differences in the data formats between different image layouts, and therefore, the PSI scores vary widely. Data in the same format have similar PSI scores due to the same image layout, similar privacy protection preferences, and tampering locations. In practice, the average of the PSI scores calculated by sampling 10 data in the same format is used instead of the PSI of all the training data.

4 Conclusion and Future Work

In this paper, we propose a federated mechanism called FedRE that can simultaneously achieve robustness and effectiveness benefits with LDP protection. It considers different privacy preferences on privacy-sensitive information of clients, perturbs the parameters adaptively, and aggregates parameters based on the distribution of perturbed information. It not only achieves better privacy protection, decreasing the similarity between the reconstructed images and raw images in sensitive regions, but also reduces the noise when aggregating and improves the performance of the model.

While FedRE presents a robust framework for federated learning with privacy preservation, there are areas that merit further exploration and improvement. In future research, PSI computational efficiency can be further enhanced by employing optimization methods like utilizing the sparsity of the Jacobian matrix and leveraging approximate computation methods. Additionally, the exploration of optimizing the sampling computation of PSI scores to approximate the overall distribution effectively, especially in scenarios characterized by high data heterogeneity and imbalance, is also a noteworthy area of investigation.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 62376103, 62302184, 62436003 and 62206102; Major Science and Technology Project of Hubei Province under grant 2024BAA008; Hubei Science and Technology Talent Service Project under grant 2024DJC078; and Ant Group through CCF-Ant Research Fund. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

- [1] Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. 2022. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 4 (2022), 1–23.
- [2] Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. 2019. Local differential privacy for deep learning. *IEEE Internet of Things Journal* 7, 7 (2019), 5827–5842.
- [3] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.
- [4] David Byrd and Antigoni Polychroniadou. 2020. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–9.
- [5] Anda Cheng, Peisong Wang, Xi Sheryl Zhang, and Jian Cheng. 2022. Differentially private federated learning with local regularization and sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10122–10131.
- [6] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*. 1655–1658.
- [7] Tianchi Deng, Yingguang Li, Xu Liu, and Lihui Wang. 2023. Federated learning-based collaborative manufacturing for complex parts. *Journal of Intelligent Manufacturing* 34, 7 (2023), 3025–3038.
- [8] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems* 33 (2020), 16937–16947.
- [9] Zishan Gu, Ke Zhang, Guangji Bai, Liang Chen, Liang Zhao, and Carl Yang. 2023. Dynamic activation of clients and parameters for federated learning over heterogeneous graphs. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 1597–1610.
- [10] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 7232–7241.
- [11] Jinwoo Jeon, Kangwook Lee, Sewoong Oh, Jungseul Ok, et al. 2021. Gradient inversion with generative image prior. *Advances in neural information processing systems* 34 (2021), 29898–29908.
- [12] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.
- [13] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The composition theorem for differential privacy. In *International conference on machine learning*. PMLR, 1376–1385.
- [14] Latif U Khan, Madyan Alsenwi, Ibrar Yaqoob, Muhammad Imran, Zhu Han, and Choong Seon Hong. 2020. Resource optimized federated learning-enabled cognitive internet of things for smart industries. *IEEE Access* 8 (2020), 168854–168864.
- [15] Jingtao Li, Adnan Siraj Rakin, Xing Chen, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. 2022. Rssfl: A resistance transfer framework for defending model inversion attack in split federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10194–10202.
- [16] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* 37, 3 (2020), 50–60.
- [17] Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. 2024. Towards Efficient Replay in Federated Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12820–12829.
- [18] Zhaohua Li, Le Wang, Guangyao Chen, Muhammad Shafq, et al. 2023. A survey of image gradient inversion against federated learning. *Authorea Preprints* (2023).
- [19] Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. 2022. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems* 64, 4 (2022), 885–917.
- [20] Ruixuan Liu, Yang Cao, Yanlin Wang, Lingjuan Lyu, Yun Chen, and Hong Chen. 2023. PrivateRec: Differentially Private Model Training and Online Serving for Federated News Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4539–4548.
- [21] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. 2020. Federated learning for open banking. In *Federated Learning: Privacy and Incentive*. Springer, 240–254.
- [22] Yaxin Luopan, Rui Han, Qinglong Zhang, Chi Harold Liu, Guoren Wang, and Lydia Y Chen. 2023. Fedknow: Federated continual learning with signature task knowledge integration at edge. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 341–354.
- [23] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.
- [24] Qiying Pan, Yifei Zhu, and Lingyang Chu. 2023. Lumos: Heterogeneity-aware federated graph learning over decentralized devices. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 1914–1926.
- [25] Zhen Qin, Shuiguang Deng, Mingyu Zhao, and Xueqiang Yan. 2023. FedAPEN: Personalized Cross-silo Federated Learning with Adaptability to Statistical Heterogeneity. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1954–1964.
- [26] Zeyu Qin, Liuyi Yao, Daoyuan Chen, Yaliang Li, Bolin Ding, and Minhao Cheng. 2023. Revisiting Personalized Federated Learning: Robustness Against Backdoor Attacks. *arXiv preprint arXiv:2302.01677* (2023).
- [27] Md Mahmudur Rahman and Sanjay Purushotham. 2023. FedPseudo: Privacy-Preserving Pseudo Value-Based Deep Learning Models for Federated Survival Analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1999–2009.
- [28] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine* 3, 1 (2020), 119.
- [29] Qiheng Sun, Xiang Li, Jiayao Zhang, Li Xiong, Weiran Liu, Jinfei Liu, Zhan Qin, and Kui Ren. 2023. Shapleyfl: Robust federated learning based on shapley value. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2096–2108.
- [30] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. 2020. LDP-Fed: Federated learning with local differential privacy. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*. 61–66.
- [31] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. 2023. DaFKD: Domain-aware Federated Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20412–20421.
- [32] Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. 2022. Protect privacy from gradient leakage attack in federated learning. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 580–589.
- [33] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, and H Vincent Poor. 2021. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing* 21, 9 (2021), 3388–3401.
- [34] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security* 15 (2020), 3454–3469.
- [35] Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. 2020. A framework for evaluating gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397* (2020).
- [36] Xidong Wu, Zhengmian Hu, Jian Pei, and Heng Huang. 2023. Serverless federated aupre optimization for multi-party collaborative imbalanced data mining. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*. 2648–2659.
- [37] Gang Yan, Hao Wang, Xu Yuan, and Jian Li. 2023. CriticalFL: A Critical Learning Periods Augmented Client Selection Framework for Efficient Federated Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2898–2907.
- [38] Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. 2019. Ffd: A federated learning based method for credit card fraud detection. In *Big Data–BigData 2019: 8th International Congress, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25–30, 2019, Proceedings 8*. Springer, 18–32.
- [39] Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. 2019. Ffd: A federated learning based method for credit card fraud detection. In *Big Data–BigData 2019: 8th International Congress, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25–30, 2019, Proceedings 8*. Springer, 18–32.
- [40] Xiyuan Yang, Wenke Huang, and Mang Ye. 2023. Dynamic Personalized Federated Learning with Adaptive Differential Privacy. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- [41] Jingwei Yi, Fangzhao Wu, Bin Zhu, Jing Yao, Zhulin Tao, Guangzhong Sun, and Xing Xie. 2023. UA-FedRec: untargeted attack on federated news recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5428–5438.
- [42] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. 2021. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16337–16346.
- [43] Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. 2020. BatchCrypt: Efficient homomorphic encryption for Cross-Silo federated learning.

- In *2020 USENIX annual technical conference (USENIX ATC 20)*. 493–506.
- [44] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. 2020. Gan enhanced membership inference: A passive local attack in federated learning. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.
- [45] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610* (2020).
- [46] Jingwen Zhao, Yunfang Chen, and Wei Zhang. 2019. Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access* 7 (2019), 48901–48911.
- [47] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. 2021. Federated meta-learning for fraudulent credit card detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 4654–4660.
- [48] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. *Advances in neural information processing systems* 32 (2019).