Privacy Protection and Security Defense of Federated Learning Big Data Based on Differential Privacy

Jinmei Li
School of Cyberspace Security
Northwestern Polytechnical University, China
ljmei@mail.nwpu.edu.cn

Abstract: *With the explosive growth of network data, data privacy protection has become a focus for many researchers. To better balance privacy protection and data utility and strengthen defense during model training, a Federated Learning (FL) method based on Differential Privacy (DP) is proposed. This method evaluates the weights uploaded by participants by constructing a scoring function. The random sampling based on an exponential mechanism is introduced to enhance data privacy. The proposed framework leverages federated learning and differential privacy to mitigate the adverse impact of attackers and untrusted parties. From the experimental results, the image quality reached the highest on the Structural Similarity Index (SSIM), at 0.7, while the multi-scale structural similarity was 0.65, demonstrating strong visualization quality. When the participant was 100, the classification accuracy reached 83.5%. In addition, the classification accuracy was 90.2% when the privacy budget was 3. The research has proven that the proposed method can effectively improve image quality and classification accuracy while protecting image data privacy, providing strong support for enterprises and units to legally and efficiently utilize data.*

Keywords: *Differential privacy, federated learning, DCT, data protection, security defense*.

Received July 09, 2025; accepted March 01, 2026
https://doi.org/10.34028/iajit/23/5/14

1. Introduction

With the digital development of enterprises and units, privacy protection surrounding data is becoming increasingly prominent. In the past, the only institutions capable of controlling citizens' personal data in large quantities were government agencies with public power. However, now many enterprises and certain individuals can also possess massive amounts of data, even surpassing government agencies in some aspects He *et al*. [8] and Cao *et al*. [4]). The powerful data storage capability possessed by the backend of big data technology systems will be repurposed after assigning different meanings to the data [2]. The "secondary use" by the technical backend often involves the unauthorized use for projects unrelated to the original purpose of data collection by the data subjects. As a result, discrimination against individuals and phenomena such as information distortion and forgery may occur during the process of data aggregation and information dissemination [5-18]. To address the aforementioned issues, existing image-processing and classification methods have begun to explore big data and artificial intelligence technologies while ensuring data security. Differential Privacy (DP), as an effective privacy protection technology, has demonstrated its advantages in various applications. In addition, Federated Learning (FL), as an emerging distributed machine learning method, allows multiple participants to jointly train models without exchanging data, providing a new solution for data privacy.

Awan *et al*. [1] proposed a big data security framework that combined FL and encryption mechanisms to address the privacy protection in the Internet of Things environment. This framework effectively addressed data leakage in distributed devices and enhanced overall security protection capabilities. To optimize the adaptability of local DP in FL, Miao *et al*. [14] constructed an adaptive compression mechanism that significantly improved communication efficiency and model accuracy while ensuring data privacy. Kerkouche *et al*. [10] theoretically and empirically demonstrated the improvement effect of compression strategies on the performance of FL models under DP. A reasonable compression mechanism could weaken the impact of noise introduction on model accuracy. Hidayat *et al*. [9] focused on resource constrained scenarios of edge devices and proposed a privacy protection FL scheme with resource adaptation capability, which ensured privacy while reducing energy consumption. Zhao *et al*. [20] analyzed the robustness of unreliable participants in federated deep learning and proposed a collaborative privacy protection mechanism, effectively reducing the interference of malicious participants on model training. Liu *et al*. [12] designed the FedSel algorithm, which introduced a Top-k dimension selection mechanism under local DP constraints, significantly improving the

stability and accuracy of model training on high-dimensional data. McMahan *et al*. [13] proposed the DP FedAvg framework, which introduced DP mechanism into the FL process and achieved user privacy protection. Dong *et al*. [6] built a DP approach to address the privacy algorithm synthesis based on divergence relaxation. The effectiveness of the developed tool was demonstrated through the improvement analysis of privacy assurance by providing noise stochastic gradient descent.

In summary, although many privacy protection methods have been proposed, there is still a gap in balancing privacy protection and data availability, and model accuracy often degrades in untrustworthy participants. A method based on DP and FL (DP-FL) is proposed to improve data availability while ensuring data privacy. The proposed DP-FL method constructs a scoring function to evaluate the quality of parameters uploaded by participants. Then, the random sampling is introduced using an exponential mechanism, aiming to make data perturbation more effective and enhance privacy protection. The innovation of the proposed DP-FL method lies in combining DP and FL, proposing a new privacy protection model that can ensure image processing and classification performance. By introducing uncertainty into parameter aggregation, the method improves the security of parameter uploading and enhances robustness against untrusted participants. The proposed method provides an efficient and privacy protection data processing approach for medical image classification tasks.

While Discrete Cosine Transform (DCT) and (DP) (DCT+DP) sanitization, robust scoring rules, and the DP exponential mechanism have been extensively explored, their interactions within a unified privacy-preserving FL workflow remain largely unexplored. The proposed DP-FL framework does not simply apply these mature components independently, but rather combines them into a tightly coupled pipeline where frequency domain sanitization, score-based update evaluation, and DP sampling operate in a mutually reinforcing manner. DCT+DP reduces the dimensionality and sensitivity of client updates, directly influencing the behavior of the scoring rules and making malicious perturbations easier to detect. The scoring function then provides structured feedback, shaping the selection probability of the exponential mechanism to achieve random yet privacy-preserving aggregation. This cross-layer interaction forms a coordinated defense path, from image-level sanitization to update-level robustness, and then to aggregation-level privacy assurance.

2. Methods and Materials

There is a data privacy breach when using data. Meanwhile, lower data availability can lead to decreased data quality and data silos [17]. To comprehensively address these challenges, a two-stage approach is integrated: taking DCT and DP noise injection for image level privacy protection to improve visual fidelity while protecting privacy. The secure and robust model training of FL is achieved, identifying and filtering unreliable participants through scoring functions and exponential mechanisms. Therefore, to enhance model security while protecting data privacy, a Privacy Protection and Security Defense method for FL Big Data Based on DP is proposed. The proposed privacy protection and security defense method compresses and protects image data through DP to improve image data quality, and adopts FL to address the data silos in model training.

2.1. Image Data Protection Method Based on DP and DCT

The proposed image protection module first applies DCT to grayscale pixel matrices and then injects DP noise into the selected coefficients. The proposed method performs DCT on the grayscale pixel matrix of the image, and uses a selection function to obtain a compression matrix with feature information. DP is used for protection processing. Although DP noise can be directly added to the original image, applying DCT before noise injection has two benefits. It concentrates signal energy into fewer low-frequency components, reducing redundancy, and improving the efficiency of noise disturbance within a fixed privacy budget. This helps to maintain image quality under strong privacy restrictions. However, high-frequency components typically possess diagnostic related features. Therefore, structural similarity measures and limited expert review by radiologists are adopted in the subsequent experiments to assess whether the compressed images preserve sufficient diagnostic information. The detailed evaluation results are reported in the Results section. DCT compresses images by reducing the high-frequency components of the image [16]. The DCT is shown in Equation (1):

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot cos(\frac{\pi}{N}(n+\frac{1}{2})k) \;\; for k = 0,1,2,\ldots,N-1 \quad (1)$$

In Equation (1), *X[k]* represents the coefficient after DCT. *X[n]* signifies the sample value of the original signal. *n* signifies the length of the signal. *k* represents the frequency index. The DCT inverse transform is displayed in Equation (2).

$$x[n] = \frac{2}{N}\sum_{k=0}^{N-1} X[k] \cdot cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right), n = 0,1,2,\ldots,N-1 \quad (2)$$

The flowchart of image compression using DCT is shown in Figure 1.

In Figure 1, DCT performs a positive transformation on the input image, dividing it into high-frequency information and low-frequency information. The energy of the image signal is concentrated in fewer low-frequency coefficients. The lower frequency part of the image mainly contains important structural and contour

information, while the higher frequency part often contains more small details and noise. During the compression process, high-frequency information often has a relatively small impact on visual perception. The high-frequency coefficients obtained after DCT are selectively discarded [11]. DP protects sensitive information by perturbing the data with calibrated noise [7, 19]. DP supports different composition modes. In the general parallel composition case illustrated in Figure 2, a dataset I can be conceptually partitioned into disjoint subsets and each subset is processed by an independent randomized algorithm. However, in the proposed federated setting, clients do not operate on disjoint subsets of the same database. On the contrary, in each round of communication, only one sanitized update is released to the server.

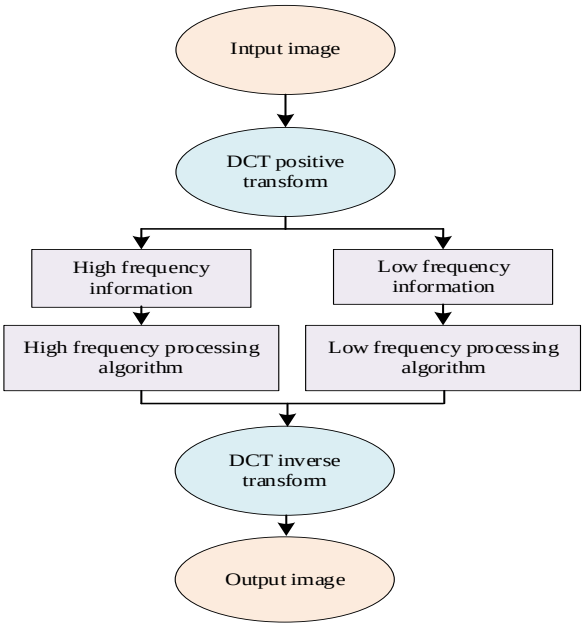


Figure 1. Flow chart of image compression using DCT.

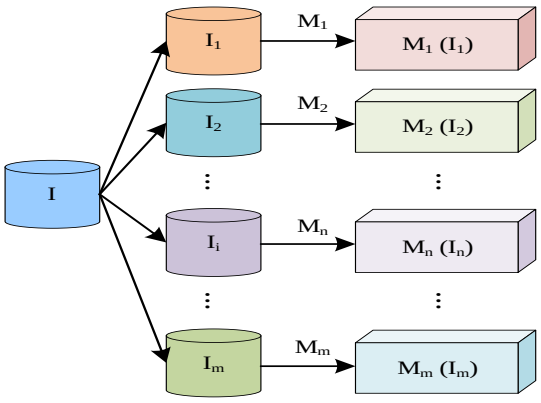


Figure 2. Schematic diagram of DP parallel combination.

As shown in Figure 2, parallel combination describes the overall situation where privacy-preserving mechanisms are applied independently to disjoint parts of the datasets. This diagram is used only to illustrate the basic idea of privacy-preserving combination. In the DP-FL framework shown in Figure 2, each client uploads a privacy update once per round. Therefore, in round T, according to the sequential combination, the total privacy budget is $\varepsilon_{total}=T_{\varepsilon}$. A satisfies ε -DP. For all adjacent datasets D and D', as well as the set of all possible output results S, Equation (3) is satisfied.

$$Pr[A(D) \in S] \leq e^{\varepsilon} \cdot Pr[A(D') \in S \tag{3}$$

In Equation (3), Pr[A(D') ∈S represents the probability that the output result belongs to set S when running algorithm, A on dataset D. ε represents the privacy budget, which measures the privacy loss of searching for individual information under a given query. For some query function f, sensitivity is defined as the maximum possible change in the output when evaluated on any such adjacent data set. Formally, the global sensitivity is given by Equation (4).

$$\Delta f = \max_{D,D} \| f(D) - f(D') \| \tag{4}$$

The norm represents the size of the output difference. This equation ensures that the added noise is calibrated to the worst-case impact that any single individual data may have on the query results.

In the proposed DP-FL framework, the query function f corresponds to the DCT-transformed and clipped model update. Before applying DP, each local update vector g_k is clipped such that $\|g_k\|_2 \leq B$. Since the DCT matrix is orthonormal, the transformation preserves the L_2-norm. For any pair of adjacent datasets, the difference between the corresponding updates is bounded by $\Delta f=\|f(D)-f(D')\|_2 \leq 2B$, which gives a concrete upper bound on the global sensitivity in the setting.

To achieve DP, noise addition is commonly used, with the most common being the Laplacian noise mechanism. Given the sensitivity Δf of query f, the generated result is displayed in Equation (5).

$$\tilde{f}(D) = f(D) + Lap(\frac{\Delta f}{\varepsilon}) \tag{5}$$

In Equation (5), *Lap* represents extracting noise from the Laplace distribution. Given the sensitivity bound, the Laplace mechanism is instantiated as $\tilde{f}(D)=f(D)+\text{Lap}(0,b)$, where the noise scale is chosen as $b=\Delta f/\varepsilon=2B/\varepsilon$. This choice follows the standard Laplace mechanism and ensures that each release of $\tilde{f}(D)$ satisfies ε-DP. Smaller ε\varepsilonε yields stronger privacy at the cost of larger perturbation, while larger ε varepsilon trades privacy for higher utility. The probability density function is displayed in Equation (6).

$$p(x \mid \mu, b) = \frac{1}{2b} exp\ (-\frac{|x-\mu|}{b}) \tag{6}$$

In Equation (6), b represents the parameter of the Laplace distribution. Based on the above theory, the proposed image protection module utilizes DCT+DP to protect image data, as shown in Figure 3.

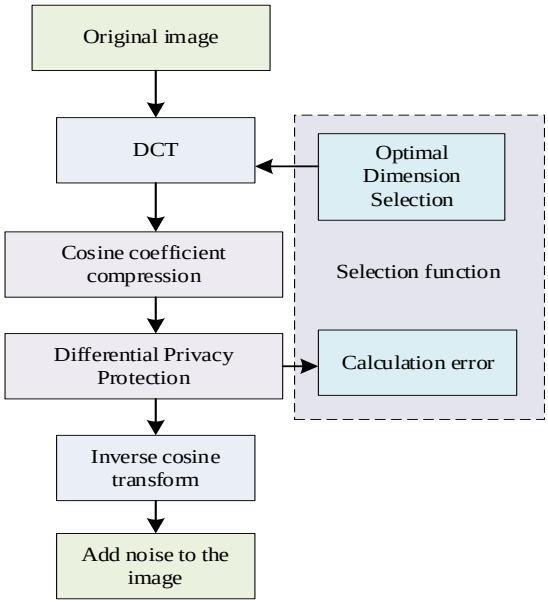


Figure 3. Framework diagram of image data privacy protection method combining DCT and DP.

In Figure 3, the original image is first transformed by DCT and compressed in the frequency domain. The resulting cosine coefficient matrix is then used as the input to the DP mechanism. Equations (7), (8) and (9) decompose the total distortion into compression error *CE(k)* and noise error *NE(k)* as functions of the retained DCT dimension k. Both errors are expressed as standard deviations in the same coefficient scale, so that the selection function *U(k)* can balance image quality and privacy perturbation under a given sensitivity bound.

$$CE(k) = \sqrt{\sum_{i=1}^{n^2-k^2} (c_i - \bar{c})^2 / n^2 - k^2} \tag{7}$$

In Equation (7), *CE(k)* represents the standard deviation of compression error. n^2-k^2 represents the number of cosine coefficients removed during the compression process. c_i represents the compressed DCT coefficient, and $\bar{c}_i$ is the mean. k represents the compression dimension of the image. The noise error is represented, as shown in Equation (8).

$$NE(k) = \sqrt{2(\Delta_1 D^{k*k}/\varepsilon)^2} = \sqrt{2}\frac{\Delta_1 D^{k*k}}{\varepsilon} \tag{8}$$

In Equation (8), $NE(k)$ represents the standard deviation of noise error. D^{k*k} represents the image compression matrix. Based on the above two types of errors, a function about the dimension k is finally obtained, as shown in Equation (9).

$$U(k) = CE(k) + NE(k) \tag{9}$$

In Equation (9), *U(k)* represents the selection function of dimension k. *CE(k)* and *NE(k)* are both computed as the standard deviation of the same unit (DCT coefficient scale), thus dimensionally consistent for addition in *U(k)*.

2.2. Model Training Method Based on Federated Learning

Based on the above method, the proposed DCT+DP module addresses the image privacy protection. To enhance the availability of data based on privacy protection, a model training method based on FL is proposed. The proposed FL training method adds local perturbation and security defense modules based on FL, aiming to enhance the defense capability against model attacks in the training model. FL is a machine learning framework where participants take encrypted private data for computation, exchanging only the parameters, weights, and gradients of the encrypted model [3, 15]. The structure of FL is shown in Figure 4.

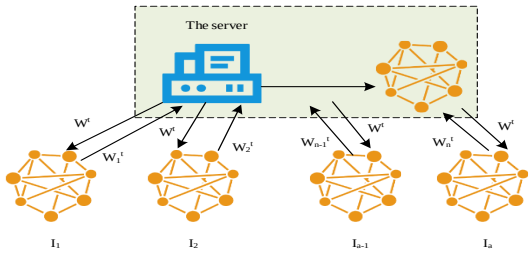


Figure 4. FL principle structure diagram.

As shown in Figure 4, each participant trains the model on local data and uploads parameters to the server. The server aggregates them into a global model and broadcasts it back to clients for iterative updates. Although FL can effectively assist in model training, there are security issues, such as the possibility of attack participants using modified model structures to disrupt the joint model. The attack method for attack participants in FL is shown in Figure 5.

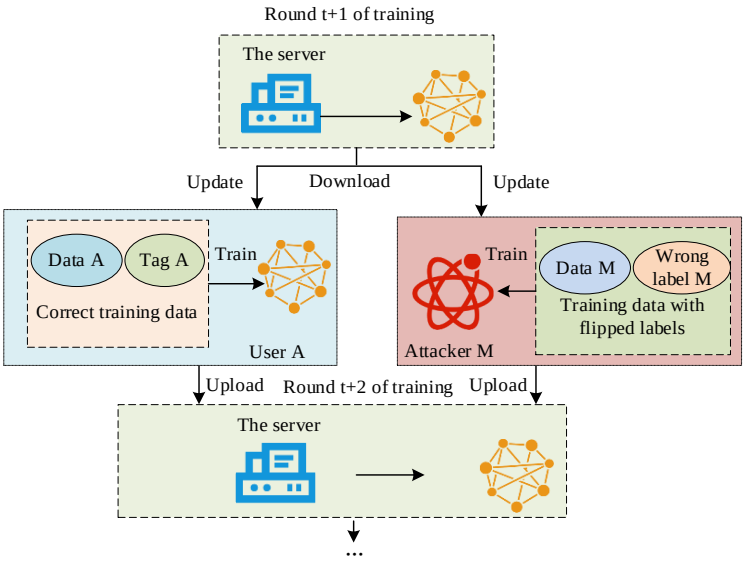


Figure 5. Attack methods of attack participants on the joint model in FL.

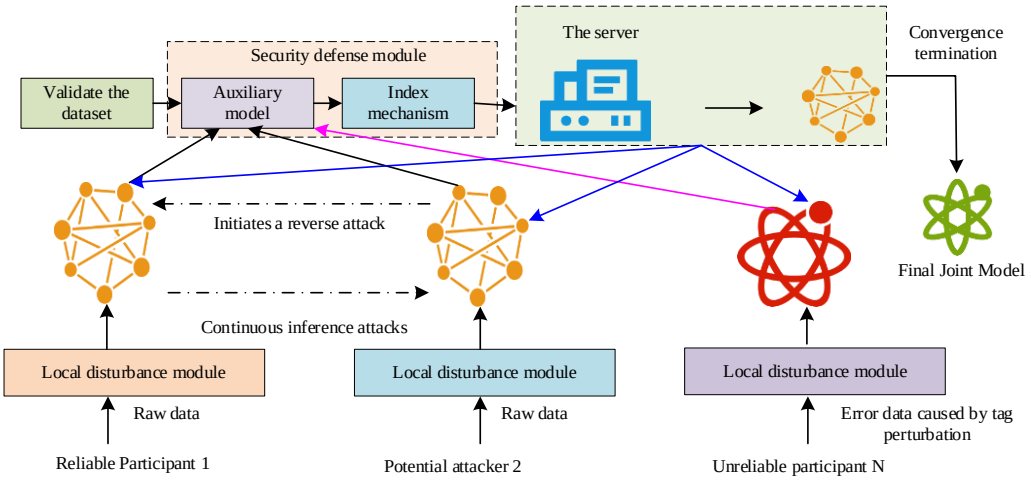


Figure 6. Framework diagram of data privacy protection and security defense system based on FL.

In Figure 5, the attack participant first updates the local model with the issued parameters, then trains the updated model using flip labels and uploads the incorrect model parameters. This type of erroneous parameter is aggregated and averaged together with other parameters on the server. After multiple rounds of communication, the joint model is severely damaged and becomes unusable. Therefore, a system based on FL is proposed. The system framework is shown in Figure 6.

In Figure 6, there are several types of participants among all parties, namely trusted party, attacker, and untrusted party. The trusted party expresses the desire to effectively communicate and cooperate with participants through the system. The attacker is a participant who aims to steal the privacy of other participants. Untrusted parties refer to participants who aim to disrupt the model. Firstly, all participants will obtain the basic model from the service provider, and then train the basic model through their own or collected data, and obtain the weight parameters of the model. All participants will transmit the obtained weights to the server. The server will calculate the average of the uploaded weight parameters based on the transmission results to obtain the final weight parameters. The obtained parameters are transferred to all participants to update their respective models and proceed to the next round of parameter

selection. When the joint model converges or reaches the preset training termination condition, the final model is transmitted to all parties. To evaluate the reliability of each participant, a score function $u(G, D_v, \tilde{W_i})$ is computed on a shared validation set D_v. In the classification task, the score is defined as Equation (10).

$$u(G, D_v, \widetilde{W}_i) = \frac{l_i}{|D_V|} \tag{10}$$

In Equation (10), u represents the score function calculated by the server for each participant. G represents the set of weight parameters. D_v represents the validation set. $\tilde{W_i}$ represents the weight uploaded by all parties. l_i represents the number of correctly predicted samples by the participants. $|D_v|$ represents the total number of samples in the validation set. According to this method, the server can evaluate the quality of parameters uploaded by participants to screen out untrusted parties. Although this form coincides with validation accuracy, it plays the utility function in the framework: It quantitatively measures the consistency between client (sanitized) updates and global validation distribution. Clients whose updates significantly deviate from the majority will obtain much lower scores and thus a much smaller probability of being selected in the subsequent exponential mechanism. Therefore, the uncertainty is incorporated into the parameters. The specific method is to randomly sample the scores of all parties through an exponential mechanism, and aggregate and average the sampled parameters, as shown in Equation (11).

$$Pr[\widetilde{W}_i \in G] \propto exp\left(\frac{\varepsilon \cdot u(G, D_v, \widetilde{W}_i)}{2 \cdot M \cdot \Delta u}\right) \tag{11}$$

In Equation (11), P_r represents the sampling probability. M signifies the number of parameters. Δu signifies the sensitivity of u. The exponential mechanism then uses the score in equation (10) to stochastically select trustworthy updates under DP constraints. Specifically, the sampling probability of participant i is defined as $P[\widetilde{W}_i \in G] \propto exp\left(\frac{\varepsilon u(G,D,\widetilde{W}_i)}{2 \cdot M \cdot \Delta u}\right)$, where M is the number of parameters and Δu is the sensitivity of the score function. In each communication round, the server first computes the score u for all received updates, then applies Equation (11) to sample a subset of clients, and finally aggregates only the selected updates. In this way, the exponential mechanism is fully integrated into the training loop as a DP-compliant and robustness-aware client selection step. On the local end, participants extract features from the local dataset to obtain a one-dimensional vector set, which is then normalized, as shown in Equation (12).

$$Z_i = \frac{x_i - \overline{x}}{\sigma} \tag{12}$$

In Equation (12), z_i represents the normalized result. x_i represents the real value in a one-dimensional vector. $\bar{x}$ is the average value of a one-dimensional vector. σ represents the standard deviation. $\bar{x}$ is shown in Equation (13).

$$\overline{x} = \frac{1}{r}\sum_{i=1}^{r} x_i \tag{13}$$

In Equation (13), r represents the length of a one-dimensional vector. σ is shown in Equation (14).

$$\sigma = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(x_i - \overline{x})^2} \tag{14}$$

The normalized value is encoded and converted into a binary value (0 represents positive, and 1 represents negative), as shown in Equation (15).

$$g(i) = ([2^{-k} \cdot |\ z\ |] mod 2)_{k=-m}^{n}, i = k + m + 1 \tag{15}$$

In Equation (15), $g(i)$ represents the converted i-th bit code. n represents the binary digits of the integer part. m represents the binary digits of the decimal part. Based on the above steps, the local training feature set can be obtained. The overall procedure of the proposed method can be summarized as follows. On each client, raw image features are extracted and arranged into a one-dimensional vector, which is normalized according to Equations (12)-(14). The normalized vector is then encoded into a binary representation as described in Equation (15) to form the local training feature set. After local model training, each client uploads a sanitized update, which is scored by Equation (10) on the server side. The server applies the exponential mechanism in Equation (11) to sample client updates according to their scores and aggregates the selected ones to update the global model. These steps are repeated over multiple communication rounds until convergence. The pseudocode for the DCT+DP privacy protection algorithm at the image end is shown in Algorithm 1.

Algorithm 1. Pseudocode for (FL) Based on (DP) Enhancement and DCT Image Protection

```
Input: Local image sets {Dk}, global model θ0, rounds T,
       DCT dimension d, clipping bound C,
       privacy budgets (ε1, ε2, and εsel)

For each client k do
    For each image x in Dk do
        X ← DCT(x)
        c ← retain_low_frequency(X, d)
        c̃ ← c + Laplace(Δ1/ε1)            // DP protection of
image features
    End For
    D̃k ← {c̃}
End For

For t = 1 ... T do
    Server broadcasts θt-1

    For each client k do
        θk ← LocalTrain(θt-1, D̃k)
        uk ← θk - θt-1
        ûk ← Clip(uk, C)
        v ← DCT(ûk)
        ṽ ← v + Laplace(Δ2/ε2)            // DP protection of
```

model update
ũk ← IDCT(ṽ)
End For

// Scoring + Exponential mechanism
For each client k do
scorek ← Validate(θt-1 + ũk)
*wk ← exp(εsel * scorek / (2Δs))*
End For
Select subset S according to wk

$\theta t \leftarrow \theta t\text{-}1 + (1/|S|) * \Sigma (\tilde{u}k \mid k \in S)$

End For

3. Output: Final global model θTResults

Due to the limited availability of real multi-institutional medical data, the NIH ChestX-ray14 public dataset is selected for simulation experiments. To approximate a realistic FL environment, the training data is divided into non-IID partitions to simulate heterogeneous client distributions, and each client performs local training before participating in parameter aggregation. Communication rounds and aggregation steps are explicitly executed to simulate the federated communication process. All participants train and communicate independently in a unified environment, still following the federated communication mechanism for parameter aggregation. To verify the effectiveness of the data privacy protection and security defense method, the Structural Similarity Index (SSIM) and Multi-Scale Structural Similarity Index (MS-SSIM) are taken as indicators to assess the image quality of the method. Meanwhile, visualized images using DCT and DP methods are presented. In addition, Random Forest (RF) is used to compare the classification accuracy of images. In terms of model training, the training effect of the FL is analyzed. The training classification accuracy of the designed method is compared with the latest similar methods. Meanwhile, the impact of parameter selection on model training is explored.

This FL system consists of a central server and multiple clients, each storing its own medical image data locally. All clients download the global model and upload local updates. Communication is assumed to be authenticated and encrypted. The threat model considers considers two adversarial scenarios: Privacy Adversarial: An honest but curious server or external observer might attempt to infer sensitive information from transmitted model updates. To mitigate this risk, the proposed DP-FL framework applies DCT compression and DP to local representations and parameter updates, providing (ε,δ)-DP guarantees. Utility Adversarial: Some clients may engage in malicious behavior while complying with communication protocols. These adversarial clients can modify labels or manipulate gradients but cannot access the original data from other clients. The adversarial experiments implement implements four typical malicious behaviors: label flipping, model poisoning, backdoor insertion, and Byzantine random updates. The server is trusted to correctly aggregate updates but cannot inspect the original data. Under this threat model, the proposed scoring and exponential mechanism aims to filter out unreliable updates, while DP noise ensures individual-level privacy is protected even if some components of the system are curious or compromised.

3.1. Analysis of image data protection methods based on differential privacy

The dataset used in the experiments is derived from the ChestX-ray14 collection, and the samples corresponding to three thoracic disease categories, pneumonia, including pulmonary edema, and tuberculosis, are selected for the experiments. All experiments are implemented in Python 3.8. The deep learning and machine learning components rely on scikit-learn 1.2.2, PyTorch 1.13.1, and NumPy 1.23.4. Experiments are conducted on Ubuntu 20.04 using an Intel i7-11700 CPU with 32 GB RAM. To ensure reproducibility, all software versions, hardware configurations, and hyperparameter settings are kept fixed across all runs. For DP processing, the Laplace mechanism is employed, where the noise scale is computed as the ratio between the empirical sensitivity and the privacy budget ε. The global sensitivity is set to 1 according to the standard DP practice. The RF classifier is configured with 100 trees and a maximum depth of 15. In the FL setup, the initial learning rate is 0.01, and exponential decay with a factor of 0.95 is applied every 10 communication rounds. All experiments are repeated five times using fixed random seeds (42, 43, 44, 45, and 46) to control randomness in data shuffling, parameter initialization, and DP noise sampling. Final results are reported as the average over the five runs. The Chest X-Ray dataset is randomly partitioned into training and testing subsets with an 80:20 split. All preprocessing steps, training configurations, and evaluation protocols are kept consistent to ensure experimental reproducibility.

Similar methods, including Sine Transform Combined with DP (SCT+DP) and Wavelet Transform combined with DP (WT+DP) are compared with the proposed DCT+DP. The visualization results are shown in Figure 7.

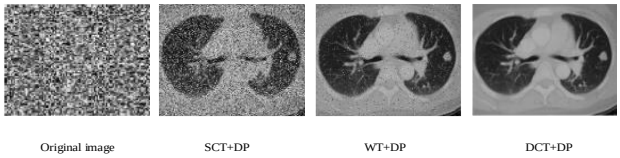
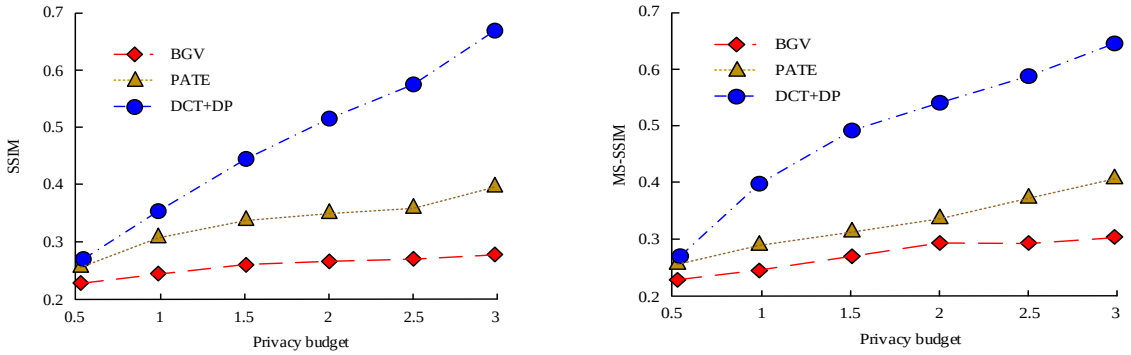


Figure 7. Comparison of the effects of different methods on image processing.

In Figure 7, after SCT+DP processing, the original image showed a significant improvement in image clarity, but there was still some noise present. After WT+DP processing, the image visualization quality was significantly improved, and the noise removal effect was

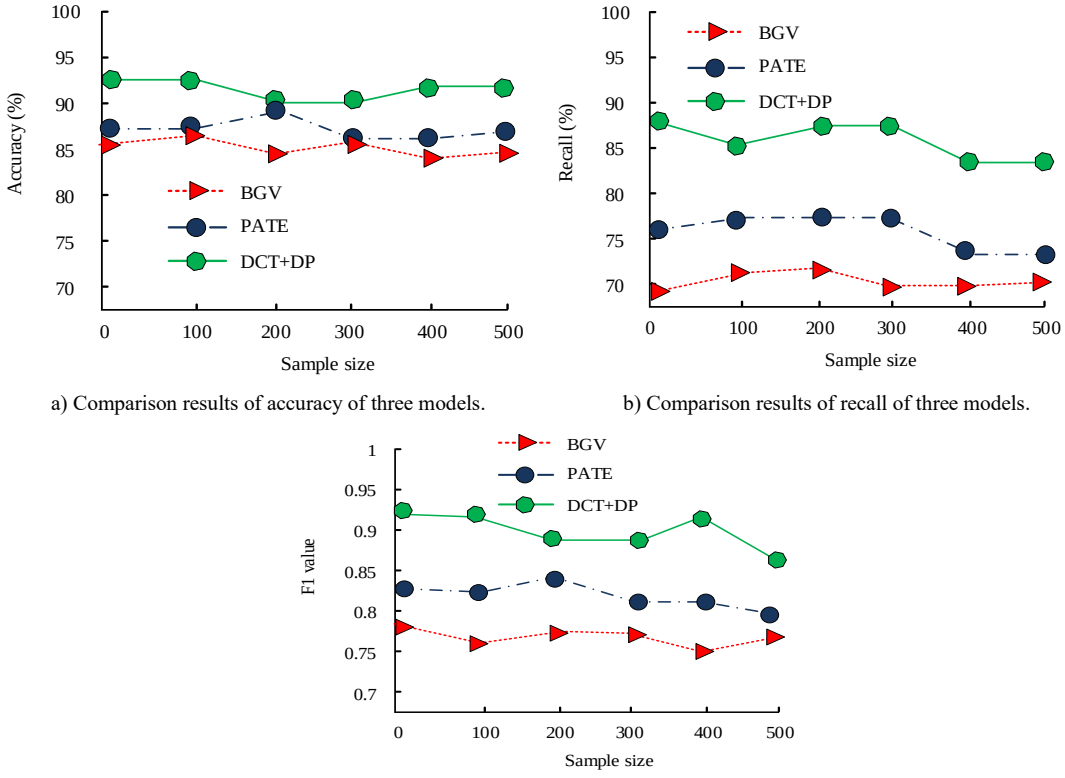
better compared with SCT+DP denoising effect. After DCT+DP processing, the image quality was significantly better compared with the other two methods. All three algorithms can improve the quality of image visualization while considering image privacy. Overall, DCT+DP achieves the best balance between image privacy and quality among comparison methods. To verify the impact of different privacy budget scenarios on image quality, Brakerski-Gentry-Vaikuntanathan (BGV) and Private Aggregation of Teacher Ensembles (PATE) are taken as comparison models. The results are shown in Figure 8.



a) SSIM evaluation results. b) SSIM evaluation results.

Figure 8. Comparison of SSIM and MS-SSIM evaluation results of different models.

In Figure 8, the privacy budget is a parameter used to measure the privacy protection, with lower values indicating higher protection. As shown in Figure 8 a), the method based on BGV showed a relatively small increase in SSIM with the increase of privacy budget. When the privacy budget reached 3, SSIM was 0.3. The SSIM value of images processed using PATE increased with the increase of privacy budget. Compared with BGV, APTE images had less distortion. The DCT+DP method has better performance compared with the other two methods, with a maximum SSIM of 0.7. According to Figure 8 b), in terms of MS-SSIM evaluation, DCT+DP had a higher MS-SSIM value compared with the other two methods. When the privacy budget was 3, the MS-SSIM value was 0.65. The DCT+DP outperforms the comparison algorithm in evaluating SSIM and MS-SSIM, indicating that the proposed method can effectively maintain image quality while protecting the privacy of image data. To further demonstrate the image quality protection of the method based on image privacy protection, RF is used to classify the dataset. The results are shown in Figure 9.



a) Comparison results of accuracy of three models. b) Comparison results of recall of three models.

c) Comparison results of F1 value of three models.

Figure 9. Comparison of classification performance across multiple model pipelines under varying sample sizes.

According to Figure 9 a), in the accuracy evaluation, the average accuracy using BGV was 86.2%, the classification accuracy using PATE was 88.4%, and the classification accuracy using the proposed DCT+DP method was 92.8%. According to Figure 9 b), in the recall evaluation, the average recall using BGV was 70.4%, the classification recall using PATE was 76.8%, and the classification recall using DCT+DP method was 88.6%. According to Figure 9 c), in the evaluation of F1 values, the F1 value using BGV was 0.78, the PATE was 0.82, and the DCT+DP method was 0.91. Experimental data show that the image quality is highest after using the DCT+DP method, which has better classification. The proposed method can ensure good usability of image data while protecting privacy. BGV, PATE, and DCT+DP represent three heterogeneous model pipelines with different feature extraction and learning mechanisms. Although they do not employ the same neural network architecture, the structural diversity they introduce is equivalent to evaluating multiple deep learning models. The consistent advantage of DCT+DP in these heterogeneous settings indicates that the proposed framework is model-independent and can be generalized to different network architectures.

3.2. Analysis of Model Training Effectiveness Based on Federated Learning

To verify the training effectiveness of the FL-based, the classification performance is evaluated under different conditions. Firstly, the classification accuracy of the model is analyzed under different numbers of participants. Meanwhile, the model classification under different numbers of untrusted parties is analyzed, as displayed in Table 1.

Table 1. Comparison of model classification structures under different number of participants.

Type	Accuracy
S=P=50, Q=0	82.2%
S=P=75, Q=0	82.6%
S=P=100, Q=0	83.5%
S=100, P=0.7S, Q=0.3S	81.7%
S=100, P=0.6S, Q=0.4S	80.5%
S=100, P=0.5S, Q=0.5S	80.2%
S=100, P=0.4S, Q=0.6S	79.9%
S=100, P=0.3S, Q=0.7S	79.5%

In Table 1, S signifies the total number of participants, P signifies the number of normal participants, and Q signifies the number of untrusted parties. According to Table 1, as the total number of participants increased, the classification accuracy gradually improved. When the total number of participants was 100, the model classification accuracy was 83.5%. This is because as the number of participants increases, the server receives more parameters that are closer to the theoretical values. As the untrusted party increases, the model classification accuracy has decreased, but it still remains at a high level. When the number of untrusted parties was 70, the model classification accuracy was 79.5%. This indicates that the proposed model has a good effect on eliminating the impact of untrusted parties and can effectively avoid the model performance degradation caused by untrusted parties. To investigate the impact of parameter M on model performance, experiments are conducted for analysis. The results are shown in Figure 10.

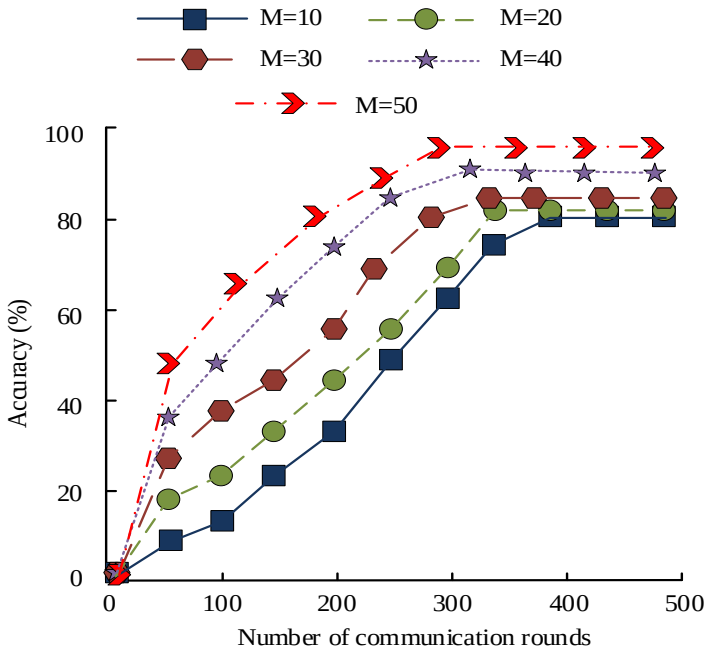
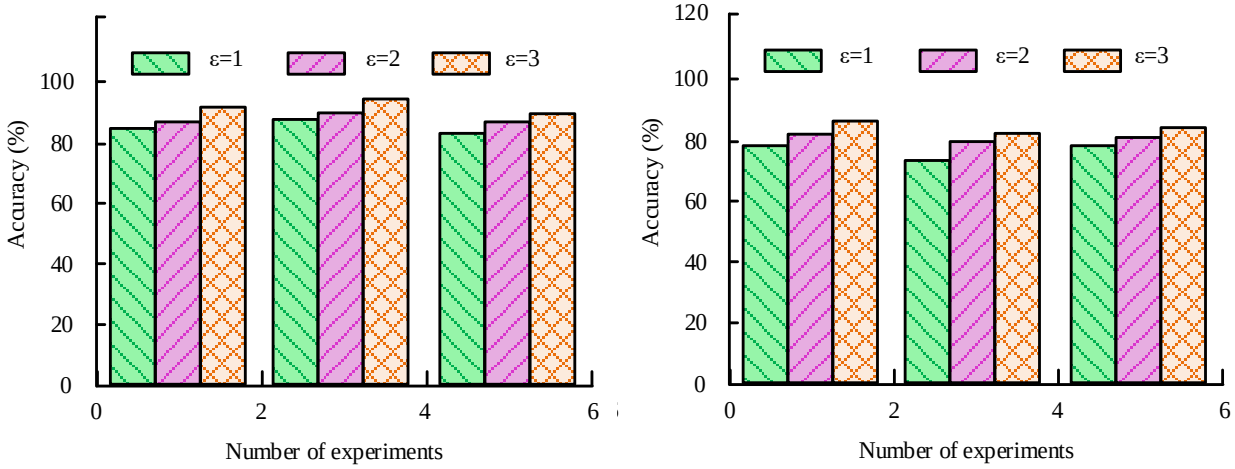


Figure 10. Comparison of model classification accuracy under different parameter numbers.

In Figure 10, M signifies the number of parameters that affect the number of participants selected by the server in each communication. When M=10, the average classification accuracy of the model was 85.4%. As M

increased, the convergence speed significantly increased, and the classification accuracy also improved to some extent. When M=20, the model converged with an accuracy of 86.5% after 380 communication rounds. When M=30, the model converged with an accuracy of 87.2% after 368 communication rounds. When M=40, the model converged with an accuracy of 88.6% after 350 communication rounds. When M=50, the model converged when the communication round reached 300, and the classification accuracy at convergence was 92.6%. Experimental data show that the size of M has obvious impacts on the convergence speed and affects its classification performance. Meanwhile, the communication cost of different M values was evaluated. As M increased from 10 to 50, the number of communication rounds decreased from 420 to 300, while the average time per round increased by approximately 9.7%. This indicates that increasing the sampling size improves convergence speed at the cost of slightly higher per-round decay, demonstrating a trade-off between efficiency and precision. This may be due to the fact that as M increases, the joint parameters obtained by the server become closer to the values obtained through joint training. The classification accuracy of the model under different privacy budgets is explored, and the results are shown in Figure 11.



a) Changes in model accuracy when Q=0

b) Changes in model accuracy when Q=0.55

Figure 11. Classification accuracy of the model under different privacy budgets.

According to Figure 11 a), when the untrusted party Q=0 and the privacy budget was 1, the average classification accuracy of the proposed model was 85.2%. With the increase of privacy budget, the image privacy protection decreased, and the image classification accuracy gradually increased. When the privacy budget was 3, the average classification accuracy was 90.2%. It should be noted that the privacy budget ε is not tuned to maximize accuracy. Instead, several representative values (e.g., 0.5, 1, 2, 3) are selected to reflect varying levels of DP, commonly adopted in GDPR-compliant scenarios. The chosen privacy budget values follow widely accepted practices in DP research and real-world deployments. In the experimental evaluation, ε is restricted to a moderate range of approximately 0.1-5, because an excessively small ε injects prohibitive noise, whereas an excessively large ε provides insufficient protection Therefore, the values ε=0.5, 1, 2, and 3 represent standard low-, medium-, and high-privacy regimes that allow meaningful evaluation of privacy-utility trade-offs. As shown in Figure 11 b), when Q=0.5S and privacy budget was 1, the average classification accuracy showed a similar growth trend as when Q=0. The average classification accuracy was lower compared with Q=0. As the privacy budget increased, the accuracy gradually increased. When the privacy budget was 3, the accuracy was 86.6%. Experimental data show that the privacy budget size has obvious impacts on the classification performance. The proposed model can maintain high classification accuracy with a lower privacy budget and has strong performance. Meanwhile, the model can effectively resist the adverse effects of untrusted participants. Moreover, since model training involves multiple communication rounds, the total privacy loss is accumulated across rounds. T denotes the number of training rounds. The total privacy cost under the standard DP composition becomes $\varepsilon_{total} = T \cdot \varepsilon$. In the configured settings, T=100 and per-round ε=0.03 are set, resulting in a cumulative privacy loss of ε_{total} =3.0. This upper bound is used for final evaluation and remains within acceptable privacy risk levels under current standards. A fixed $\delta = 10^{-5}$ is used to constrain the failure probability. However, SSIM or MS-SSIM alone cannot reflect true privacy guarantees. Therefore, the empirical privacy loss (ε,δ)-DP level is reported. $\delta = 10^{-5}$ is selected as a standard privacy leakage tolerance, following existing benchmarks such as those used in GDPR-compliant systems. The effective ε values are derived based on the total number of queries and the sensitivity of the cosine coefficient matrix.

From the perspective of full DP composition accounting, the proposed method can be decomposed into two types of mechanisms. The first is a one-shot

image-level mechanism M_1 (DCT+DP) applied locally on each client before model training. For each client, M_1 satisfies (ε_1, δ_1)-DP with respect to its own medical images. Since different clients hold disjoint datasets, these image-level mechanisms compose in parallel across clients, so the overall privacy cost of M_1 is still bounded by (ε_1, δ_1) for each client. The second part is an iterative parameter-update mechanism M_2 run during FL. In each communication round t=1, …, T, the server-side aggregation combined with local perturbation induces a mechanism M_t that satisfies (ε_2, δ_2)-DP. Based on standard sequential composition, the training process satisfies ($T\cdot\varepsilon 2$, $T\cdot\delta 2$)-DP with respect to each client's contribution to the model parameters.

Combining the preprocessing and training stages, the end-to-end system satisfies ($\varepsilon_1+T\cdot\varepsilon 2$, $\delta_1+T\cdot\delta_2$)-DP for each client. In implementation, the image-level DCT+DP mechanism is instantiated once per image, while the federated training runs T=100 rounds with a per-round privacy budget ε_2=0.03 and a fixed $\delta_2=10^{-5}$. This yields an upper bound of approximately ε_2, total=3.0 for the training phase. Since ε_1 is of the same order of magnitude but applied only once and δ is dominated by the strictest guarantee, a total privacy budget on the order of $\varepsilon_{total}\approx 3.0$ with $\delta=10^{-5}$ for each client is conservatively reported, which is consistent with commonly accepted settings in DP-based medical data analysis.

As shown in Figure 12, a privacy-utility curve is plotted to demonstrate the trade-off between image classification accuracy and privacy budget.

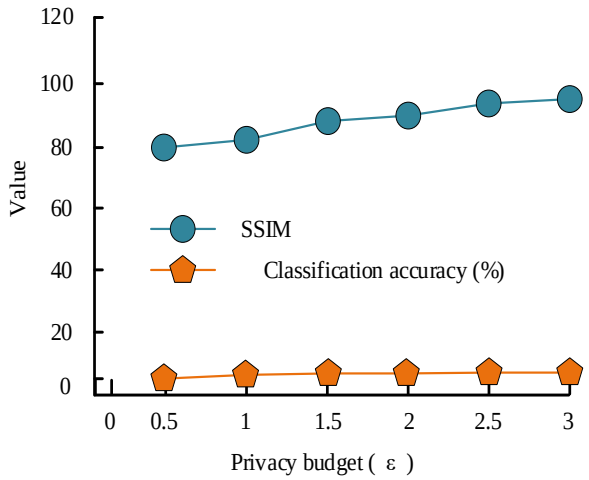


Figure 12. Privacy-utility trade-off curve under differential privacy (δ=1e-5)

As shown in Figure 12, as ε increased from 0.5 to 3.0, the intensity of noise injection weakened and the image distortion decreased. Therefore, SSIM steadily increased from 0.51 to 0.70, and the image quality improved significantly. The accuracy of image classification also increased from 80.3% to 90.2%, demonstrating better model performance. The accuracy of different aggregation rules under test standard adversarial FL attacks is shown in Table 2.

Table 2. Test accuracy of different aggregation rules under adversarial FL attacks.

Aggregation method	Label-flip (%)	Model poisoning (%)	Backdoor (%)	Byzantine (%)
FedAvg	78.4	72.9	69.3	70.8
Trimmed mean	83.9	81.5	79.8	80.6
Krum	84.6	82.1	80.7	81.4
Bulyan	85.3	83.4	81.5	82.2
Proposed (Scoring+EM)	86.7	84.9	83.1	83.7

Table 2 compared the performance of different aggregation rules under four representative adversarial FL attacks. Overall, FedAvg showed the lowest robustness across all scenarios, with test accuracy dropping to 78.4% under label-flip and further to 69.3% under backdoor injection. This indicates that simple averaging cannot effectively mitigate malicious parameter updates.

For the statistically robust methods, Trimmed Mean consistently improved accuracy by 4%-10% compared with FedAvg, achieving 83.9% under label-flip and 79.8% under backdoor attacks. Krum further enhanced robustness by filtering outliers based on distance, reaching 84.6%-82.1% across attacks. Bulyan demonstrated the strongest resilience among classical robust aggregators by combining multi-stage outlier removal and median-based selection, improving accuracy to 85.3% under label-flip and 82.2% under Byzantine attacks. The proposed method obtained the highest accuracy in all adversarial settings, reaching 86.7% under label-flip and maintaining 83.1% under backdoor attacks.

The improvement on Bulyan is moderate (1.2-1.6%), demonstrating that incorporating uncertainty-based parameter sampling can further suppress malicious contributions while preserving benign updates. These results confirm that the proposed scoring-based selection combined with the exponential mechanism yields stronger robustness compared with both non-robust and classical robust aggregation rules. The performance of different FL frameworks and privacy protection baselines in benign environments is shown in Table 3.

Table 3. Performance of different FL frameworks and privacy protection baselines in benign environments.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 value
FedAvg	88.3	87.5	84.1	0.858
FedProx	88.9	88.0	84.7	0.863
FedNova	89.4	88.7	85.5	0.870
DP-SGD	87.1	86.2	83.2	0.846
DP-FedAvg	89.0	88.2	85.1	0.868
PATE-CTGAN	88.1	87.3	83.9	0.855
Proposed DP-FL	90.6	89.8	86.9	0.883

Table 3 compared multiple FL frameworks and privacy-preserving baselines under a benign (non-adversarial) environment. Among classical FL

algorithms, FedAvg yielded an accuracy of 88.3%, while FedProx and FedNova provided moderate improvements (88.9% and 89.4%, respectively), reflecting their enhanced stability in heterogeneous data distributions.

Privacy protection variants show different levels of utility reduction: DP-SGD exhibited the largest decrease (87.1%) due to per-step gradient noise, while DP-FedAvg maintained higher performance (89.0%) by aggregating privatized local updates rather than privatizing every gradient computation. PATE-CTGAN achieved an accuracy of 88.1%, indicating that synthetic-data-based privacy mechanisms can preserve reasonable utility but still fall short compared with DP-based approaches.

The proposed DP-FL framework obtained the highest performance across all metrics, achieving 90.6% accuracy and an F1 value of 0.883. These results demonstrate that the integration of DCT-based dimensionality reduction, DP perturbation, and scoring-exponential mechanism sampling improves both model generalization and privacy-utility balance, outperforming existing FL and DP baselines. The ablation experiments and statistical significance of the key components are shown in Table 4.

Table 4. Ablation experiments and statistical significance of the key components (five runs).

Variant	SSIM	MS-SSIM	Accuracy (%)	F1 value	*p*-value
Full model	0.700±0.012	0.652±0.015	90.2±0.6	0.882	/
w/o DCT compression	0.624±0.018	0.571±0.021	86.5±0.9	0.844	0.004
w/o DP noise (no privacy)	0.732±0.014	0.691±0.017	91.1±0.7	0.891	0.031
w/o scoring function	0.691±0.013	0.641±0.016	87.8±0.8	0.861	0.012
w/o exponential mechanism	0.683±0.016	0.636±0.018	88.5±0.7	0.867	0.019
DP-FedAvg baseline	0.664±0.017	0.612±0.019	88.9±0.8	0.872	0.027

Table 5 shows that removing any key component leads to a measurable performance drop. Eliminating DCT compression results in the largest decrease in image quality (SSIM decreased from 0.700 to 0.624) and reduced accuracy to 86.5%, confirming its importance for preserving structural information under privacy constraints. Removing DP noise slightly increases accuracy (91.1%) but violates privacy. The small *p*-value indicates that the utility gain is statistically significant yet modest. Excluding the scoring function or the exponential mechanism lowers accuracy to 87.8% and 88.5%, respectively, demonstrating their contribution to filtering unreliable updates and improving aggregation robustness. Compared with the DP-FedAvg baseline, the full model consistently achieves higher SSIM, MS-SSIM, and F1 value. Overall, the ablation confirms that each module meaningfully enhances image quality, privacy protection, or model stability.

4. Discussion and Conclusions

With the increasing emphasis on data privacy protection, it has become increasingly important for enterprises and units to effectively utilize data without leaking information. The proposed DP-FL method evaluated and aggregated the parameters uploaded by participants. The privacy protection was introduced by combining random noise to ensure the security of local data. The proposed DCT+DP achieved the best balance between image privacy and quality, and maximized image quality without compromising data privacy. Compared with the two comparison methods, DCT+DP had a higher MS-SSIM value. When the privacy budget was 3, the MS-SSIM value was 0.65. This indicates that the proposed DCT+DP can maintain high image quality when the image protection is average, and performs better compared with similar methods. The classification performance of the same model largely depended on image quality. The F1 value of the model using BGV was 0.78, the F1 value of the PATE was 0.82, and the DCT+DP method was 0.91. According to the data, the image quality using the DCT+DP method was higher compared with current advanced methods, which could enhance the classification performance. The proposed DP-FL method uses exponential mechanism sampling to probabilistically weight unreliable parameter updates, thereby reducing the impact of attackers and untrusted parties, although further comparison with established robust aggregation rules is required. When the number of untrusted parties was 70, the classification accuracy was 79.5%, still maintaining a high classification accuracy. The proposed method successfully establishes a balance between privacy protection and classification effectiveness, and has significant practical value. Although the proposed DP-FL method has achieved promising results, there may still be challenges when dealing with more complex or large-scale datasets. Future work can focus on improving the adaptability of the model to complex datasets and further optimizing privacy budget adjustment strategies.

References

[1] Awan K., Din I., Almogren A., and Rodrigues J., "Privacy-Preserving Big Data Security for Iot with Federated Learning and Cryptography," *IEEE Access*, vol. 11, pp. 120918-120934. 2023. DOI: 10.1109/ACCESS.2023.3328310

[2] Bandara F., Jayawickrama U., Subasinghage M., Olan F, and et al., "Enhancing ERP Responsiveness Through Big Data Technologies:

An Empirical Investigation" *Information Systems Frontiers*, vol. 26, no. 1, pp. 251-275, 2024. https://doi.org/10.1007/s10796-023-10374-w

[3] Beikmohammadi A., Khirirat S., and Magnusson S., "On the Convergence of Federated Learning Algorithms Without Data Similarity," *IEEE Transactions on Big Data*, vol. 11, no. 2, pp. 659-668, 2024. DOI: 10.1109/TBDATA.2024.3423693

[4] Cao Y., Wei W., and Zhou J., "Privacy Protection Data Mining Algorithm in Blockchain Based on Decision Tree Classification," *Web Intelligence*, vol. 20, no. 2, pp. 103-112, 2022. https://doi.org/10.3233/WEB-210485

[5] Demirol D., Das R., and Hanbay D., "A Key Review on Security and Privacy of Big Data: Issues, Challenges, and Future Research Directions," *Signal, Image and Video Processing*, vol. 17, no. 4, pp. 1335-1343, 2023. https://doi.org/10.1007/s11760-022-02341-w

[6] Dong J., Roth A., and Su W J., "Gaussian differential privacy," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 1, pp. 3-37, 2022. https://doi.org/10.1111/rssb.12454

[7] Goncalves C., Bessa R., and Pinson P., "Privacy-Preserving Distributed Learning for Renewable Energy Forecasting," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 3, pp. 1777-1787, 2021. DOI: 10.1109/TSTE.2021.3065117

[8] He Y., Ouyang W., Li S., Wang L., and et al., "Cloud Computing Data Privacy Protection Method Based on Blockchain," *International Journal of Grid and Utility Computing*, vol. 14, no. 5, pp. 480-492, 2023. https://doi.org/10.1504/ijguc.2023.133436

[9] Hidayat M., Nakamura Y., and Arakawa Y., "Privacy-Preserving Federated Learning with Resource-Adaptive Compression for Edge Devices," *IEEE Internet of Things Journal*, vol. 11, no. 8, pp. 13180-13198, 2024. DOI: 10.1109/JIOT.2023.3347552

[10] Kerkouche R., Acs G., Castelluccia C., Geneves P., "Compression Boosts Differentially Private Federated Learning," *in Proceedings of the IEEE European Symposium on Security and Privacy*, Vienna, pp. 304-318, 2021. DOI: 10.1109/EuroSP51992.2021.00029

[11] Koley S., Roy H., Dhar S., and Bhattacharjee D., "Cross-Modal Face Recognition with Illumination-Invariant Local Discrete Cosine Transform Binary Pattern (LDCTBP)," *Pattern Analysis and Applications*, vol. 26, no. 3, pp. 847-859, 2023. https://doi.org/10.1007/s10044-023-01139-x

[12] Liu R., Cao Y., Yoshikawa M., and Chen H., "FedSel: Federated SGD under Local Differential Privacy with Top-k Dimension Selection," *in Proceedings of the 25th International Conference on Database Systems for Advanced Applications*, Jeju, pp. 485-501, 2020. https://link.springer.com/chapter/10.1007/978-3-030-59410-7_33

[13] McMahan H., Ramage D., Talwar K., and Zhang L., "Learning Differentially Private Recurrent Language Models," *arXiv preprint*, arXiv:1710.06963v3, pp. 1-14, 2017. https://doi.org/10.48550/arXiv.1710.06963

[14] Miao Y., Xie R., Li X., Liu X., and et al, "Compressed Federated Learning Based on Adaptive Local Differential Privacy," *in Proceedings of the 38th Annual Computer Security Applications Conference*, Texas, pp. 159-170, 2022. https://doi.org/10.1145/3564625.3567973

[15] Pei J., Liu W., Li J., Wang L., and Liu C., "A Review of Federated Learning Methods in Heterogeneous Scenarios," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5983-5999. 2024. DOI: 10.1109/TCE.2024.3385440

[16] Shao Z., Wang X., Tang Y., and Shang Y., "Trinion Discrete Cosine Transform with Application to Color Image Encryption," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 14633-14659, 2023. https://doi.org/10.1007/s11042-022-13898-6

[17] Thakkar O., Ramaswamy S., Mathews R., and Beaufays F., "Understanding Unintended Memorization in Federated Learning," *arXiv preprint*, arXiv:2006.07490v1, pp. 1-14, 2020. https://doi.org/10.48550/arXiv.2006.07490

[18] Tunc-Abubakar T., Kalkan A., and Abubakar A., "Impact of Big Data Usage on Product and Process Innovation: The Role of Data Diagnosticity," *Kybernetes*, vol. 52, no. 9, pp. 3178-3196, 2023. https://doi.org/10.1108/K-11-2021-1138

[19] Yanamala A., "Secure and Private AI: Implementing Advanced Data Protection Techniques in Machine Learning Models," *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence*, vol. 14, no. 1, pp. 105-132, 2023. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=aivJDSwAAAAJ&citation_for_view=aivJDSwAAAAJ:qjMakFHDy7sC

[20] Zhao L., Wang Q., Zou Q., Zhang Y., and Chen Y., "Privacy-Preserving Collaborative Deep Learning with Unreliable Participants," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1486-1500, 2020. DOI: 10.1109/TIFS.2019.2939713

Jinmei Li is a senior engineer and a doctoral student at the School of Cybersecurity, Northwestern Polytechnical University. Her main research focus is on Network Information Security.