Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 6357750, 20 pages
https://doi.org/10.1155/2023/6357750

Research Article

IPCADP-Equalizer: An Improved Multibalance Privacy Preservation Scheme against Backdoor Attacks in Federated Learning

Wenjuan Lian,[1] Yichi Zhang,[1] Xin Chen,[1] Bin Jia,[1,2] and Xiaosong Zhang[2]

[1]*College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China*
[2]*Center for Cyber Security, College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China*

Correspondence should be addressed to Bin Jia; jiabin@sdust.edu.cn

Received 6 March 2023; Revised 25 July 2023; Accepted 25 August 2023; Published 7 September 2023

Academic Editor: Mohammad R. Khosravi

Although there are some protection mechanisms in federated learning, its training process is still vulnerable to some powerful attacks, such as invisible backdoor attacks. Existing research work focuses more on how to prevent attacks in distributed training scenarios and improve the security of the FL training process, but it lacks consideration of utility and robustness, especially when the learning model of FL suffers from stealth backdoor attacks. This paper proposes an improved FL defense scheme IPCADP based on user-level differential privacy and variational autoencoders technology. The scheme can control and protect the privacy attribute of the image and can also eliminate the triggers that exist in the poisoned image. The experimental results show that compared with some existing defense schemes, IPCADP can defend against invisible backdoor attacks and improve the classification accuracy of the main task, while mitigating the impact of attacks on model robustness and stability. To a certain extent, the balance and unity of security, utility, and robustness are realized.

1. Introduction

Nowadays, an increasing number of learning systems require a large amount of user privacy information during training. Therefore, to meet the needs of user data privacy protection, federated learning systems have emerged. Federated learning [1] is a distributed machine learning technology, also known as collaborative learning. It was first proposed by Google in 2017, which enables large-scale training on devices that generate data. Sensitive data is only retained by its owners, meaning local collection and local training. The model parameters are uploaded to the central training coordinator for aggregation, and a round of model updates is completed [2]. Compared with the traditional centralized data training method, this learning method keeps sensitive data locally, which provides a certain degree of data privacy protection for users participating in the training [3], but there are also security risks. Because of the local training characteristics of the client, this process cannot be controlled by the server [4], which also provides a great opportunity for malicious clients to implement backdoor attacks. The backdoor attack implants the backdoor trigger into the model by tampering with the user dataset and then induces the model to make mistakes when implementing label classification [5]. To sum up, it is precisely because there are still many security problems in federated learning that research on data privacy protection and model security in the training process of federated learning is particularly necessary.

There have been some defenses against backdoor attacks in previous work. The defense method in [6] mainly weakens or even eliminates the influence of poisoning and backdoor attacks through data enhancement, but this method is not applicable to FL, because the resources and data volume consumed by all clients in the process of data enhancement are too large, which exceeds the privacy budget. The method

proposed in [7] is also not suitable for FL, because the method proposed in this article is based on the premise that the data of each client is independent and identically distributed. However, the data held by each client of the real FL is different, which is fundamentally impossible to achieve. Considering the privacy and security requirements of FL, the scheme mentioned in [8] is not suitable for FL. It requires the defender to be able to access the local private training data of all clients while implementing defensive measures, which is completely contrary to the theoretical basis of FL.

It can be seen from the previous research that most of the traditional and effective backdoor attack-defense methods are no longer applicable in FL, and it is urgent to find new and targeted defense methods. In research in recent years, differential privacy technology has gradually been used to achieve defense against backdoor attacks. The basic differential privacy technology is divided into two levels, namely, record level [9] and user level [10, 11]. These two defense techniques start with the definition of DP to measure and control the added noise. In addition, some scholars have proposed the concept of weak differential privacy [12], which resists backdoor attacks by adding slight noise to the aggregated update, but the effect is not satisfactory due to the small magnitude of the noise. Besides, there are LDP and CDP [13]. These two schemes add a lot of noise before and after uploading parameters, respectively. Their defense effect is stronger, but this also comes at a relatively large price. While defending against attacks, they will greatly reduce the accuracy rate, which is not worth the loss. Some scholars have proposed an image enhancement strategy to replace the original DP method [14] realized the resistance to the model flipping attack by controlling the brightness, color, contrast, balance, and other attributes of the picture. This provides a way of thinking for the safe processing of images in this paper, but this method cannot modify the attributes of the things in the pictures themselves, and there are certain limitations. The CND method proposed in [15] improves the accuracy of the classification task by improving the DP-SGD algorithm and dynamically controlling the injected noise, but reduces the injected noise accordingly. This reduces the degree of protection of FL in safety, making it impossible to have both the properties of these two models.

In order to solve the above problems, in this paper, we propose a comprehensive defense backdoor attack scheme in the face recognition scenario; the scheme name is IPCADP. The specific contributions of this paper are as follows:

(1) In order to solve the problem of the significant decline in utility in the DP scheme, this paper improves on the general ULDP and restricts and adjusts the clipping threshold necessary in the model update process to a certain extent. This adjustment not only strictly complies with the requirements of user-level privacy but can also continuously adapt and change as the model is updated. Since the injected noise is proportional to the threshold, the limitation of the noise is also achieved. This paper refers to the improved scheme as AULDP.

(2) Considering the balance between privacy and utility, this paper uses VAE technology [16] to remove sensitive privacy attributes in images based on the statistics and distribution of image attributes, which are often irrelevant to the main task. At the same time, before the model is trained, the VAE technology can also be used to clean the triggers implanted in the client data by the backdoor attack, so as to prevent the damage of the poisoned data to the security of the overall model and realize the supplement of security and boost.

(3) In the experiments in this paper, we assume that the attacker can make changes to the training data of the malicious client and can also affect its training process. In the previous work [17], a backdoor attack algorithm based on BadNets was proposed, and the defense measures for this type of attack have been relatively perfect. In this paper, we leverage a newer and more powerful ISSBA stealth backdoor attack [18] to evaluate our scheme. The trigger for this attack is more difficult to catch and eliminate, the attack is more subtle, and the attack effect is more obvious.

(4) Considering the privacy, security, utility, and robustness of FL comprehensively, based on the above two schemes, this paper proposes a comprehensive defense scheme called IPCADP. And through ablation experiments and comparative experiments with other advanced methods, it is jointly confirmed that the scheme in this paper has indeed improved the accuracy of the main task, so that the utility of the model can be maintained. At the same time, it can also ensure the security and robustness of the model while reducing noise injection, reducing the success rate of brute force attacks, and maintaining the stability of the model.

The rest of this article is organized as follows: In Section 2, we will introduce detailed knowledge about FL, DP, and variational autoencoders, including but not limited to basic definitions, application formulas, specific classifications, etc. In Section 3, this paper first introduces the threat model, points out the security problems of FL, aims at these security problems, proposes our improvement scheme, and describes and demonstrates the relevant framework and details in detail. The experimental part of this article is placed in the fourth and fifth sections. First, the software and hardware settings related to the experiment are explained. Then, the ablation experiment and the comparative experiment are carried out to verify the advanced nature of the scheme from three aspects: effectiveness, security, and robustness. These aspects are discussed in detail. Finally, the conclusions are described in Section 6.

2. Preliminary Knowledge

2.1. Federated Learning. The main purpose of federated learning is to use data stored in distributed data centers as comprehensively and easily as possible. The main idea of

federated learning proposed by Google is to no longer require data sharing between the central server and each participating client, but to learn collaboratively with each other and eventually become a federation [19]. The general federated learning process is as follows: First, the central server initializes the model parameters $W_1^k, W_2^k, W_3^k, \ldots W_n^k$ and distributes the parameters to the local n clients in turn. At this time, each client uses the local dataset and the received parameters W^k to train locally and then perform local updates: $U \coloneqq W_i^k - W^k$. At this point, each client uploads the locally updated model parameters, and the server aggregates and averages them. Then, the central server uses the gradient descent algorithm to update the global model: $W^{k+1} = W^k - \eta U^{'}$, where η is the learning rate and $U^{'} = 1/n \sum U^i$. After getting the updated global model, the server distributes it again and repeats the above process until the model converges [20]. The schematic diagram of the federated learning model of the client-server architecture is shown in Figure 1:

2.2. Differential Privacy. Differential privacy technology is a kind of information-fuzzy mechanism [21]. Its main function is to add noise of different volumes to records or data and then disturb the data, so that a certain piece of data loses its uniqueness and can be hidden among a large amount of data. It is convenient to hide the sensitive information contained in the records or data to avoid leakage.

In 2006, the differential privacy protection technology was proposed by Proserpio [22], and its basic mathematical definition is as follows:

Definition 1. For any algorithm M, let any subset of its output be Ω. If the output of algorithm M on any adjacent datasets Da and $\mathrm{Da}^{'}$ satisfies the following conditions, it means that the algorithm M provides (ε, δ)-differential privacy protection:

$$\Pr[M(\mathrm{Da}) \in \Omega] \leq \exp(\varepsilon)\Pr\left[M\left(\mathrm{Da}^{'}\right) \in \Omega\right] + \delta. \quad (1)$$

Among them, the nonnegative parameter ε is the privacy budget, indicating the degree of privacy protection. The smaller its value is, the higher the degree of protection is, and the less information the algorithm M may leak. δ is also a nonnegative parameter, which represents the probability that the difference between the output results of the algorithm M on the dataset Da and $\mathrm{Da}^{'}$ exceeds exp (ε). Obviously, the smaller δ is, the higher the degree of privacy protection is.

Generally speaking, depending on the period of adding noise, differential privacy techniques are basically divided into two categories, namely, central differential privacy (CDP) and local differential privacy (LDP). In this paper, we mainly introduce local differential privacy (LDP) in detail.

2.2.1. LDP. When the third-party server that collects data is untrustworthy, centralized differential privacy technology cannot protect the privacy of local data, and local differential privacy technology needs to be used. Local differential privacy means that each client participating in federated learning first performs security processing by adding noise so that the privacy of each record or parameter can be protected locally before uploading the locally updated parameters. The definition of local differential privacy is shown in Definition 2:

Definition 2. Let P: $2^x \longrightarrow Y$, $\varepsilon > 0$, $\delta \in [0, 1)$, and algorithm P satisfies (ε, δ)-local differential privacy if and only if for all adjacent datasets D, $D^{'} \in \mathrm{X}$ and all $y \in Y$, the following inequalities hold:

$$\Pr[Q(D) \in Y] \leq \exp(\varepsilon)\Pr\left[Q\left(D^{'}\right) \in Y\right] + \delta. \quad (2)$$

Among them, 2^x is the set composed of all subsets of X, Y is the value range of the algorithm Q; and the definitions and value ranges of ε and δ are the same as Definition 1.

It can be seen from the definition that local differential privacy is to protect user privacy by ensuring the similarity of the output of any two adjacent datasets, so that the process of privacy protection is transferred from the server for data collection to the user's locale, thereby avoiding leaks during parameter collection. In LDP technology, ULDP is more strict and practical, and ULDP will be introduced in detail below.

2.2.2. User-Level Differential Privacy. As a privacy protection technology, differential privacy can naturally provide different levels of differential privacy for different scenarios and different needs. According to different levels, privacy guarantees can be mainly divided into two types, namely, record-level differential privacy [9] and user-level differential privacy [10, 11]. In this paper, we mainly introduce and utilize user-level differential privacy techniques, and we follow related works in [23, 24] for parameter setting and the introduction of relevant knowledge.

First of all, based on the relevant knowledge of DP introduced earlier, we choose to use the Gaussian mechanism with L2 norm sensitivity as the algorithm Q. In the previous related work, the general method was to perturb the relevant output $p(x)$ by adding Gaussian noise conforming to the Gaussian distribution, that is, adding Gaussian noise with a mean value of 0 and a variance of σ^2, as shown in the following equation:

$$Q(x) = p(x) + N\left(0, \sigma^2\right). \quad (3)$$

At the same time, we define the model update function in federated learning as L (D^p, θ). In general, sensitivity is defined as an upper bound on noise perturbations that satisfy local differential privacy requirements; that is, sensitivity is the maximum value of noise added to the parameters.

Given two adjacent datasets Da_k^p and $\mathrm{Da}_k^{'p}$ and the gradient $g(\mathrm{Da}_k^p) = L(\mathrm{Da}_k^p, \theta^t)$, Da_k^p indicates the local private training dataset of the kth client, t represents the tth round of global training, and θ denotes the model parameters. The

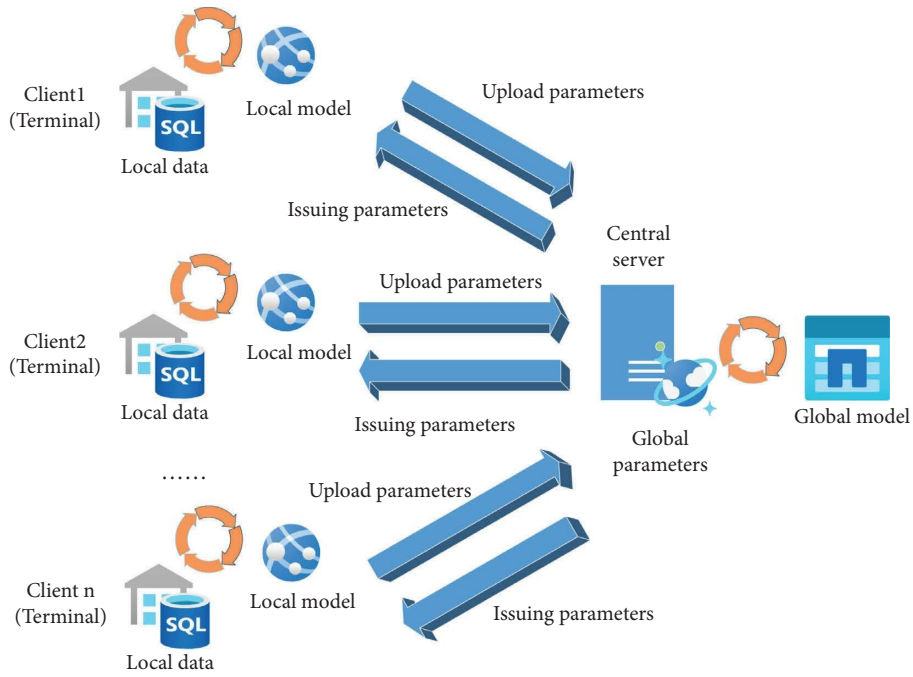

Figure 1: Client-server architecture of FL.

local maximum sensitivity of the L2 norm associated with this process is defined as follows:

$$\nabla L = \max_{\mathrm{Da}_k^p, \mathrm{Da}_k^{'p} \in X} \left\| g\left(\mathrm{Da}_k^p\right) - g\left(\mathrm{Da}_k^{'p}\right) \right\|_2. \tag{4}$$

The L2 norm sensitivity here plays an important role in two aspects. On the one hand, it is to prove the differential privacy of the Gaussian noise added in this paper. On the other hand, it is a necessary condition for calculating Gaussian noise. If the sensitivity is lacking, it cannot complete the calculation of Gaussian noise. Sensitivity is an extremely important parameter for differential privacy.

In previous work, norm clipping techniques were often used to limit the aforementioned L2 norm local sensitivity. In this paper, the norm-limited threshold is denoted as C. The smaller the threshold, the smaller the sensitivity, and the smaller the added noise. At this time, the sensitivity is limited to

$$\nabla L \leq \frac{2\eta C}{\left|\mathrm{Da}_k^p\right|}. \tag{5}$$

At the same time, given the number of global communications T, the number K of clients participating in global training each time, and the client sample ratio, that is, the ratio of clients participating in training each time $\mathrm{ra} = K/U$, where U is the total number of clients. Given ε_k and δ_k, according to the work in [20], the following inequalities can be obtained:

$$\ln \frac{1}{\delta_k} < \frac{\varepsilon_k^2 \sigma_k^2}{2T * \mathrm{ra} * \nabla L^2}. \tag{6}$$

After finishing the above formula (6), the following equation can be finally obtained. This equation is used to determine the variance σ_k of Gaussian noise satisfying (ε, δ) local differential privacy requirements of the Kth client:

$$\sigma_k = \frac{\nabla L \sqrt{2T * \mathrm{ra} * \ln\left(1/\delta_k\right)}}{\varepsilon_k}. \tag{7}$$

In formula (7), the variance σ_k of each client can be determined, and thus the corresponding Gaussian noise can also be determined. Using the above formula, the unique Gaussian noise can be determined for each client.

To sum up, the ULDP framework can design different and unique Gaussian noises for each client participating in the training and add Gaussian noises to the parameters before the client uploads the parameters, which can guarantee the training to a certain extent. The specific related algorithm process is shown in Algorithm 1.

2.3. Variational Autoencoders. Variational autoencoders, namely, VAE was first proposed by Kingma and Welling [25]. Previous articles and work pointed out that the core idea of VAE is to use an autoregressive process to train and sample latent variables and finally, obtain a highly structured training data probability distribution. The latent variable z represents the internal structure of the data x, and at the same time, z also satisfies some specific posterior distribution $p(x, z)$. It can be seen that if you want to use VAE technology to process pictures, you first need VAE to compress high-dimensional pictures into low-dimensional latent variables and then perform autoregressive-related processing on latent variables to maximize the quality of

Input: The model parameter θ^0, initial limit threshold C, Client sample proportion q, Total communication rounds T, Parameters of each client corresponding to LDP $(\varepsilon_i, \delta_i)$, Attenuation coefficient γ
Output: Global model θ
for each round $t = 0, \ldots, T\text{-}1$ **do**
  Select clients CL from U
  for each client k in CL **do**
    $g_k^t(\mathrm{Da}_k^p) \longleftarrow L(\mathrm{Da}_k^p, \theta^t)$
    $g_k^t(\mathrm{Da}_k^p) \longleftarrow g_k^t(\mathrm{Da}_k^p)/\max\ (1, \|g_k^t(\mathrm{Da}_k^p)\|_2/c)$
    $C \longleftarrow$ Create New Norm (C^t, γ, t)
    $\theta_k^{t+1} \longleftarrow \theta^t - \eta g_k^t(\mathrm{Da}_k^p)$
    $\sigma_k = \nabla L\sqrt{2T * \mathrm{ra} * \ln(1/\delta_k)}/\varepsilon_k$
    $\theta_k^{t+1} \longleftarrow \theta_k^{t+1} + N(0, \sigma_k)$
  end
  $\theta^{t+1} \longleftarrow 1/|CL| \sum_{k=1}^{CL} \theta_k^{t+1}$
end

Algorithm 1: Normal user-level differential algorithm.

pictures. The following formula (8) shows the optimization objective of VAE:

$$L\left(\theta, \psi; x^{(i)}\right) = E_{q_\psi(z|x^{(i)})}\left[\log p_\theta\left(x^{(i)}|z\right)\right] - R_{\mathrm{KL}}\left(q_\psi\left(z|x^{(i)}\right)\|p_\theta(z)\right). \tag{8}$$

Among them, θ and ψ are the model-related parameters of the encoder and decoder in VAE, respectively, and $x^{(i)}$ is a piece of data in the related dataset. Both $p_\theta(x^{(i)}|z)$ and $p_\theta(z)$ can be calculated by the Bayesian formula. $E_{q_\psi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)]$ in formula (8) is the reconstruction loss, and $R_{\mathrm{KL}}(q_\psi(z|x^{(i)})\|p_\theta(z))$ is the KL regular term. As shown in formula (9), for VAE technology, the addition of regularization items helps VAE to shape a latent space with a good structure, and at the same time, it can also alleviate the overfitting of the model on the training dataset:

$$\mathrm{KL}(p(x)\|q(x)) = \sum p(x)\log\frac{p(x)}{q(x)}. \tag{9}$$

3. Privacy Protection Scheme Based on Image Attribute Control and AULDP

3.1. Threat Model. In the setting of our threat model, we always assume that the central server is honest and that the attacker only controls some clients, which we call malicious clients or poisoned clients. A poisoned client in a threat model means that the client's local training dataset is added by an attacker to an adversarial trigger, which can manipulate the output of the model. Models trained with poisoned datasets will make arbitrary or targeted mispredictions on other data embedded with the same triggers, which are basically the results specified by the attacker. The relevant backdoor attack process utilized in this article is as follows:

First, the attacker participates in the training process of federated learning as a client, uses the method of generating poisoned data in the previous work [26] to pollute the local data of the poisoned client, and pollutes the original clean training data D_{clean} into poisoned training data D_{poison} with related triggers. Second, use both poisoned data and clean data to simultaneously train the local model of the poisoned client. The specific principle is shown in the following formula:

$$w = \mathrm{argmax}E\left[\sum_{x', y' \in D_{\mathrm{poison}}} p\left(G_t\left(x'\right) = y'\right) + \sum_{x, y \in D_{\mathrm{clean}}} p(G_t(x) = y)\right]. \tag{10}$$

Among them, w is the local model parameter, G_t is the global model issued by the server, which the client participating in the training uses as the local model, x', y' are the poisoned data and the corresponding label, and x, y are clean data and corresponding labels. The main purpose of the formula (10) is to obtain a suitable model parameter w, which can maximize the probability that poisoned data and clean data can be accurately classified as corresponding labels. Finally, the attacker uploads the trained model parameter w containing the backdoor attack vulnerability to the server, thereby polluting the global model. So far, the backdoor attack on federated learning has been realized.

In this paper, we use the ISSBA attack proposed in previous work [18], which is a stealth backdoor attack. The backdoor triggers of ordinary backdoor attacks are agnostic; that is to say, all poisoned samples contain the same trigger, so although the poisoned samples are not the same, the defender can use the same behavior to detect and handle

triggers. The ISSBA attack generates different and specific triggers for each sample, which is more difficult to defend against than general backdoor attacks. If it can be defended and the defense effect is good, it is enough to show the security of our scheme.

In the previous related research work, various security measures were relatively independent, and it was difficult to realize the protection of federated learning from multiple perspectives. To solve this problem, this paper proposes a comprehensive improvement scheme based on VAE technology, transfer learning, and AULDP. The overall process of AULDP and migration learning in this scheme and the VAE face attribute modification scheme are described in detail in Figure 2. The relevant details of the AULDP part are shown in Figure 3, and the relevant details of the VAE technology are shown in Figures 4 and 5.

3.2. Advanced User-Level DP. The improved ULDP strategy proposed in this paper is improved and enhanced on the basis of the user-level differential privacy proposed in [23] and the CND scheme proposed in [27]. The AULDP policy restricts and adjusts the pruning threshold of the model update to a certain extent, and with the continuous update of the model, the pruning threshold is continuously updated and reduced, becoming more in line with the definition and requirements of user-level differential privacy, more targeted, more specific, and more flexible to meet the needs of privacy protection.

Specifically, the improvement scheme we propose refers to first initializing a norm clipping threshold C^0 before the model is updated in each round and then combining the threshold with the same initialized model parameters. By constantly updating and adjusting the clipping threshold of the model update, the noise injected by DP can be flexibly limited so as to achieve the goal of improving the accuracy of the main task while maintaining the utility and robustness of the model.

The process of implementing threshold iterative change is described in Algorithm 2. The initialized norm clipping threshold is C^0, but as the model is updated, C^{t} is also updated accordingly, and it is the same as the previous threshold C^0 for comparison. If the new threshold is smaller than C^0, the updated threshold will be used for clipping. If it is larger than C^0, then the original threshold C^0 will be used to clip the model update. For the specific update process, one thing needs to be noted, that is, the clipping threshold is slightly adjusted in each round; that is, the attenuation coefficient γ is used to achieve the change, and the attenuation coefficient here is set to 0.99. But at the same time, it is necessary to perform special calculation operations on the threshold in some specific rounds. For example, when the number of rounds is less than 5, or when the number of rounds is a multiple of 60, the average update norm and Gaussian of each client are used. Noise is used to update the calculation of the threshold and then perform judgment and replacement operations.

Among them, M represents the number of clients participating in each round of federated learning, σ_{avg} represents the average value of σ_k for each client, which is more in line with the privacy requirements of ULDP, and client_i^{t+1} represents each updated norm of a user participating in training. For each client participating in the training, the related algorithm of the user's norm update is shown in Algorithm 3.

Algorithm 3 describes the update process of the local model norm for each client. In this paper, the number of rounds of local training of federated learning is set to 3 times, and the threshold C^t and model parameters θ^t are used to control each round. The local training process of the model calculates and updates the norm, and finally returns the updated norm.

In summary, the AULDP method proposed in this paper is jointly explained by the abovementioned algorithms 1, 2, and 3. The algorithm framework and algorithm details of the AULDP algorithm are marked and explained in detail in Figure 3. In the whole process of the algorithm, the local model update of each client participating in the training is related to the relevant clipping before uploading. After several rounds of automatic calculation and clipping adjustments, the threshold we get gradually falls into a relatively reasonable and correct range. At this time, there is no need to make large adjustments, and only need to make small adjustments within a certain interval and range.

3.3. VAE Face Attribute Modification and Protection. Generally speaking, VAE is a generative network sequence comparable to GAN [28]. We can input a z in a low-dimensional space and map it to real data in a high-dimensional space. Specifically, the dataset in this paper uses face images, so the role of VAE is to generate a face image by randomly inputting an n-dimensional vector. Then at this time, the input n-dimensional vector represents the n invisible factors that can ultimately determine the appearance of the face. For each of the n invisible factors, a corresponding distribution is generated, and sampling is performed from these distribution relationships, then a deep network can be used to finally generate the face image we need. Therefore, using VAE technology, we only need to use limited data input, and through the adjustment of hidden parameters, we can obtain an almost unlimited number of face pictures, and even many of these face images have never appeared before. The main reason why these pictures can be generated is the encoder in VAE. When operating and adjusting each privacy parameter, the encoder does not just generate a fixed number, but generates a corresponding confidence interval. This confidence Interval is a continuous value range and expression. In this way, if we sample again, we can obtain a lot of data that was not obtained before.

The overall architecture of the scheme is based on the general assumption that attackers cannot modify the execution process of the scheme, that is to say, the server has encapsulated the scheme. The client is only responsible for

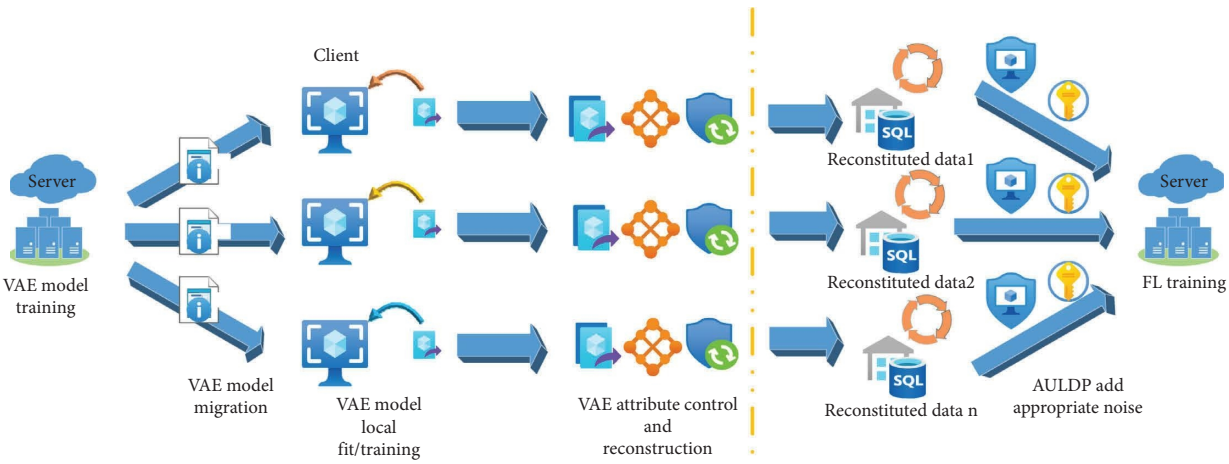

Figure 2: IPCADP overall solution architecture.

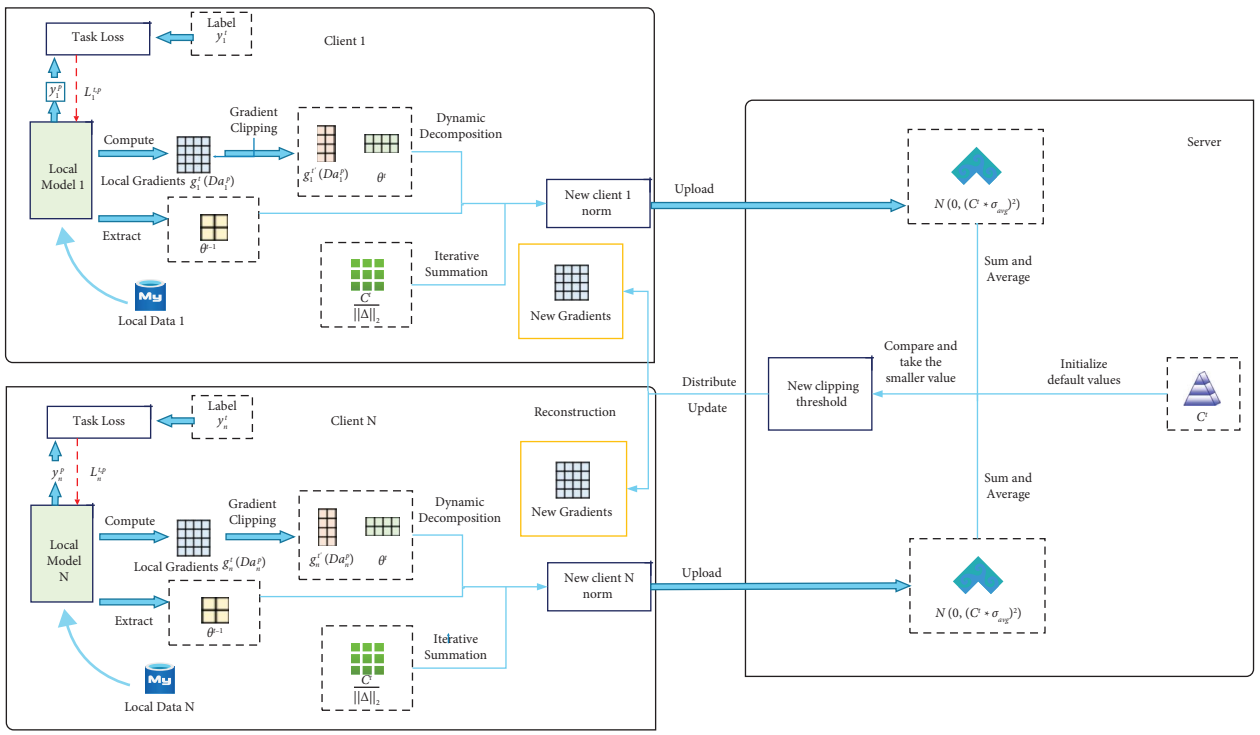

Figure 3: The specific framework and details of the AULDP program.

executing it as an application after receiving it. The overall process is as follows: First of all, in the initial stage of FL, in order to save resources on the client side, first use the collected face dataset to perform multiple rounds of training on the server side to train a VAE model. Then, with the help of some transfer learning knowledge and technology, the VAE model trained on the server side is migrated and deployed to the client. At this time, the client only needs to use the local private dataset to adapt the VAE model for several rounds. The transfer of the source domain to the target domain can be realized through pertinence training. Finally, use the abovementioned VAE model and local data labels after local adaptation on the client side, and the attribute screening method mentioned in the article [14] to separate the nonprimary attributes that can be screened and controlled, and then realize the protection of private data. For example, modify the facial expression in the image, the degree of curvature of the hair, and the color of the hair and other related attributes.

3.3.1. VAE Model Pretraining and Migration. In the basic architecture and design of FL, the computing resources and data volume held by each client participating in the training are limited. In addition, if each client trains the VAE model independently, the training results will be uneven. Because of insufficient computing resources, the training time will be too long, and it will even occupy and affect the initialization

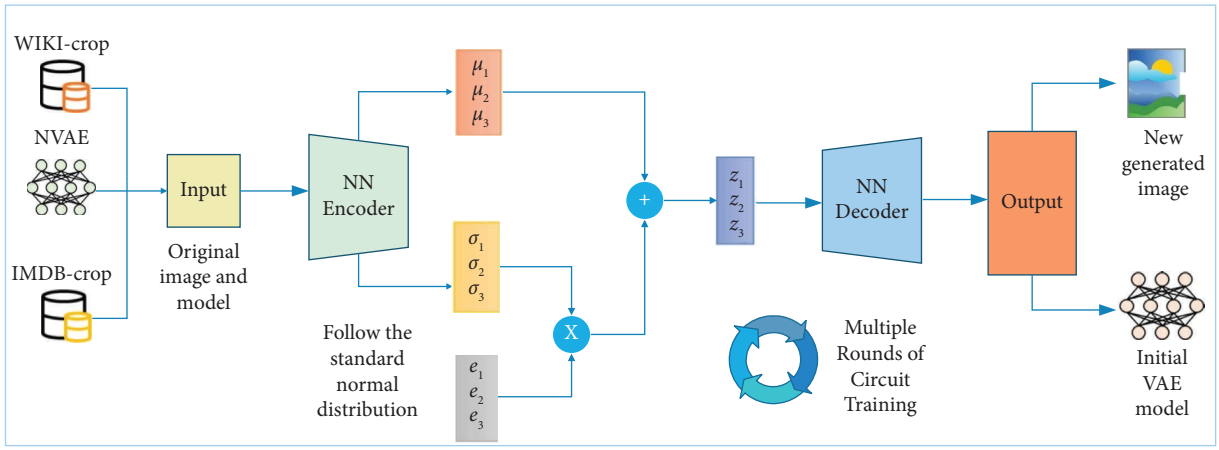

Figure 4: VAE model pretraining process details.

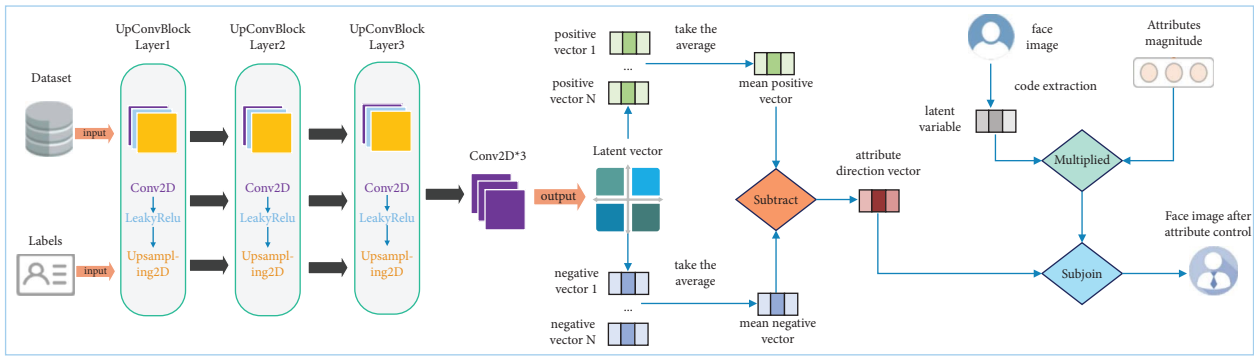

Figure 5: VAE attribute separation and attribute control process.

(1) **function** Create New Norm (C^t, γ, t)
(2) $C^{t+1} = C^t * \gamma$
(3) if $t < 5$ or $t = 60, 120, \ldots, 300$ then:
(4) $c \longleftarrow 1/M \sum_{i=1}^{M} \|\text{client}_i^{t+1}\|_2 + N(0, (C^t * \sigma_{\text{avg}})^2)$
(5) if $C < C^{t+1}$ then:
(6) $C^{t+1} \longleftarrow C$
(7) return C^{t+1}

Algorithm 2: Update of limit thresholds.

(1) **function** Client Norm Update (k, θ^t, C^t)
(2) $\theta \longleftarrow \theta^t$
(3) for each local epoch $i = 1, 2, 3$ do
(4) $\theta \longleftarrow \theta - \eta g_k^t(D_k^p)$
(5) $\Delta \longleftarrow \theta - \theta^t$
(6) $\theta \longleftarrow \theta^t + \Delta \min(1, C^t/\|\Delta\|_2)$
(7) newnorm $\longleftarrow \theta - \theta^t$
(8) return newnorm

Algorithm 3: Update of model norm.

and training time of federated learning, and affect the final model generation results. To sum up, it is not feasible to only use the client to realize the training of the VAE model, and a server with higher computing power and more abundant computing resources is needed to assist in the joint training. But this will also lead to another problem. In the basic setting of FL learning, the server does not have a training dataset, and there is no model that can perform data processing. Therefore, this paper uses VAE as the initial data processing model on the server side. Collect public-related datasets on the Internet and use them as training data on the server side. Due to the characteristics of the VAE training model, the label information in the dataset is not needed. In this article, what we need is face images to achieve gender classification, so the server only needs to collect relevant face images. The details are shown in Figure 4. Finally, after completing the training and generation of the VAE model on the server side, the finetune method in transfer learning is used to migrate the VAE model to the client and perform several rounds of adaptive training. The specific principle is shown as follows:

$$W_{D_k} = \operatorname{argmin} \sum_{x \in D_k} \operatorname{dist}\left(M_{D'}(x), x\right). \tag{11}$$

Among them, W_{D_k} is the parameter of the data processing module of the kth client, D_k is the dataset of the kth client, and $M_{D'}$ is the VAE model trained by the server. The ultimate goal of this formula is to obtain the most suitable and high-quality module parameters under the premise that the difference value distance between the face image generated by the VAE model and the face image in the real dataset x of the client is as small as possible.

3.3.2. VAE Attribute Separation. To use VAE technology to modify the attributes of face images, it is first necessary to screen and separate suitable nonmain attributes. The main reason for this is to ensure that the accuracy of the classification task will not drop significantly. In this paper, the attribute separation algorithm based on image attribute statistics and distribution rules [14] is used to realize the separation operation of nonprimary attributes.

First, use the following formula (12) to select nonprimary attributes that meet the constraints:

$$\begin{cases} \left| \sum_{y_i \in D_k} I\left(\text{lab}_{i,A_j} = \text{pos}\right) - \sum_{x \in D_k} I\left(\text{lab}_{i,A_j} = \text{neg}\right) \right| < \gamma, \\ \left| \sum_{y_i \in D_k} I\left(\text{lab}_{i,A_j} = \text{pos}, \text{lab}_{i,A_m} = \text{pos}\right) - \sum_{y_i \in D_k} I\left(\text{lab}_{i,A_j} = \text{pos}, \text{lab}_{i,A_m} = \text{neg}\right) \right| < \iota, \\ \left| \sum_{y_i \in D_k} I\left(\text{lab}_{i,A_j} = \text{neg}, \text{lab}_{i,A_m} = \text{pos}\right) - \sum_{y_i \in D_k} I\left(\text{lab}_{i,A_j} = \text{neg}, \text{lab}_{i,A_m} = \text{neg}\right) \right| < \tau. \end{cases} \tag{12}$$

Among them, A_m is the main attribute required and used by the federated learning face gender classification task in this paper, A_j is any nonmain attribute, and lab_{i,A_j} represents the i-th data of the client dataset. The attribute is the label value of A_j, pos and neg represent positive samples and negative samples, respectively, and γ, ι, and τ represent thresholds. The smaller the threshold, the better. The function of function I is to take the value 1 when the expression in it is true, otherwise it takes the value 0. At this time, it can be obtained from formula (12) that when there are only two types of labels in the sample, and a nonkey attribute is evenly distributed in the sample at this time, then this nonkey attribute can be selected as the attribute to be modified. Moreover, formula (12) can also be extended to the case of multiclassification.

In the above, we mentioned that VAE can transform the input data into highly structured latent variables, as shown in the following formula (13), encode all positive samples $y_{A_j} = \text{pos}$ and all negative samples $y_{A_j} = \text{neg}$ respectively and superimpose them into two vectors, then the difference of the vectors at this time is the attribute vector V_{A_j} that can be separated.

$$V_{A_j} = \sum_{x, y_{A_j} \in D_k} \text{encode}\left(x, y_{A_j} = \text{pos}\right) - \sum_{x, y_{A_j} \in D_k} \text{encode}\left(x, y_{A_j} = \text{neg}\right). \tag{13}$$

3.3.3. VAE Attribute Control. After obtaining the attribute vector V_{A_j} of the nonmain attribute that can be separated, according to the steps of the face attribute modification scheme, at this time, the attribute vector should be modified and controlled for the local private data of each client, and then the private Data Protection. Here, we use latent variables and formula (14) to calculate and get the modified data:

$$x_{s_A} = \text{decode}\left(\text{encode}(x) + \sum_{a \in s_A} \beta V_a\right). \tag{14}$$

First, the data x are input into the VAE model, encoded by the encoder, and then the encoded result is correlated with the attribute vector V_a in the attribute set $S_{a,y_{A_m}}$. Finally, use the decoder in the VAE model to decode the calculated latent variables, and then you can get the data x_{s_A} with modified attributes. In the process of calculating latent variables, there is a parameter β ($-1 \le \beta \le 1$), which is used to affect the performance of related data. The performance results are different when the value is positive and negative. For example, when the attribute to be controlled and changed is a smile in human facial expressions, the value of β is regular to indicate a smile, and the larger the value, the higher the degree of the smile; Conversely, a negative value of β means that the face is expressionless, and the value of β at this time can control the degree of change of facial smile expression. The flowchart of the solution to realize VAE attribute separation and attribute control is shown in Figure 5:

4. Experiments and Analysis

4.1. Datasets and Experimental Configuration. In this paper, we use the CelebA [29] and IMDB_crop, Wiki_crop [30] datasets as the training data for the federated learning face recognition gender classification task. Among them, the CelebA dataset was compiled and opened by the Chinese University of Hong Kong. It contains many face pictures and has been marked. The specific introduction is shown in Table 1. Each face picture in this dataset contains 40 attribute tags, which include whether to wear glasses, whether it is high cheekbones and other related decorative and facial features, as well as hair color, whether it is straight hair and gender and other features. This dataset is a relatively useful and authoritative dataset for face-related training, which is sufficient to meet the needs of this experiment. In this paper, we use it as the training and testing dataset of the clients participating in FL. The specific division of the training set, test set and verification set is shown in Table 1. The IMDB-crop and Wiki-crop face datasets come from IMDB and Wikipedia respectively, that is, the Internet Movie Database, which is a large database about movies, including face images, gender and age. The relevant data information in the dataset is also shown in Table 1. In this paper, we use the face images in IMDB-crop and Wiki-crop as the dataset used for server-side training of the VAE model in FL, and the split ratio is the same as above. For the convenience of experiment and display, we changed the size of the picture to 64×64, and the related face image is shown in Figures 6 and 7. The FashionMNIST dataset is a clothing and hat recognition image dataset, which includes ten different categories of items such as T-shirts, pants, and hoodies, and these image files are all $28 * 28$ grayscale images.

The CPU of the server used in the experiment is set to Intel(R) Core^{TM}i9-9960X CPU @ 3.10 GHz, the GPU used is two NVIDIA RTX 2080Ti graphics cards, the version of PyTorch is 1.11, and the version of CUDA environment is CUDA 10.0. This greatly improves the effect and efficiency of the VAE model and federated learning training.

In federated learning training for dataset CelebA and IMDB-Wiki, this paper uses ResNet-18 [31] as FL classification task model, and use Adam as the optimizer for client-side training, while on the server side, we use the FedAvg algorithm [32] as the optimizer. In the setting of this article, we assume that there are a total of 40 federated learning clients. In each round of training, 20 clients are randomly selected as the clients participating in the training in this round, and the poisoned clients are selected from each round. Choose from the specified clients, and select at most 8 clients as poisoned clients. In this paper, we select 0–8 poisoned clients in turn to test the effect of the stealth backdoor attack and the defense effect of our proposed scheme. At the same time, considering the limited computing resources and computing power of the client, we set the total number of communication rounds of FL to 300 rounds, and the number of local training rounds for each client is set to 3 rounds. For the gender classification task in this paper, the batch size is set to 32, the ε privacy budget is set to 5, the γ decay coefficient is set to 0.99, and the C^0 initial clipping threshold is set to 3.

In the federated learning training for dataset Fashion-MNIST, this paper uses MLP as the classification task model of FL, and uses the FedAvg algorithm on the server side for aggregation and update. In the setting of FL, we assume that there are a total of 20 federated learning clients, and in each round, 10 are randomly selected as the clients participating in the training in this round. The poisoned client is then selected from the selected clients in each round, and at most 4 clients are selected as poisoned clients. In this experiment, we sequentially select 0–4 poisoned clients to test the effect of stealth backdoor attack and the defense effect of our proposed scheme. Due to the limitation of computing resources and computing power of the client, we set the total number of communication rounds of FL to 100 rounds, and the number of local training rounds of each client is set to 5 rounds, the batch size is set to 128, and the ε privacy budget is set to is 10, the γ attenuation coefficient is set to 0.01, and the C^0 initial clipping threshold is set to 3.

4.2. Privacy and Security Analysis of IPCADP Method

4.2.1. AULDP Privacy Protection Analysis. This paper mainly uses noise-added privacy protection technology, that is, the advanced user-level differential privacy technology to achieve privacy protection. The differential privacy method can flexibly limit the noise injected by DP by continuously updating and accurately adjusting the pruning threshold of

Table 1: Dataset information.

Dataset	Data distribution	Number of features	Categories	Number of samples			
				Total	Train	Test	Validation
CelebA	IID	40	10177	202599	141819	40519	20261
IMDB-crop	IID	16	20284	461871	323309	92374	46188
Wiki-crop	IID	7	62328	62359	43651	12471	6237
FMNIST	Non-IID	7	10	70000	60000	6667	3333

Figure 6: IMDB-Wiki face image.

Figure 7: CelebA face image.

model update, and then add appropriate noise proportional to the pruning threshold. At the same time, this paper also employs an inductive approach to hide some sensitive attributes of the participants, obfuscating the data until the third-party cannot distinguish individuals through differential attacks. In this way, even if the attacker obtains the interactive data, he cannot infer the original data, so that the data cannot be restored, and the attacker cannot obtain the original data of each participant, thereby achieving the purpose of protecting user privacy. In the basic definition of differential privacy, there is a parameter of privacy budget ε, and the smaller the parameter, the higher the degree of privacy protection. This paper uses the AULDP method to continuously test and find the most suitable privacy budget value. At this time, the method can maintain or even improve the accuracy of the main task while ensuring a certain degree of privacy protection.

4.2.2. VAE Privacy Protection Analysis. The image attribute modification scheme in the IPCADP method can successfully realize the attribute separation and attribute control of the image. The original image is changed through VAE technology, and the sensitive privacy attributes in the image (these attributes are often irrelevant to the main task) are modified. Remove or replace to generate a reconstructed image, thereby protecting the main attributes of the image and achieving the purpose of protecting client data privacy. Not only that, this method can also eliminate and clean the triggers of the poisoned image, so as to prevent the damage of the poisoned data to the security of the overall model, and realize the supplement and improvement of security. The specific implementation and related effects are as follows.

As mentioned earlier in this article, in FL, the computing resources and capabilities of the client are much smaller than that of the server, so the task of training the VAE model is placed on the server, and after the training is completed, the model is migrated back to the client for several rounds adaptive training. Based on the NVAE model proposed in [33] on the server side, this paper uses more than 500,000 face images in IMDB-crop and Wiki-crop for 400 rounds of training to obtain a VAE model with good results. Figures 8 and 9 are the original images and the generated images of the VAE model. From the two face images, we can see that the face image generated by the trained VAE model is similar to the overall outline and basic facial features. The original images are basically the same, but there are differences in some details, but this difference does not affect the classification of gender perceptually. To sum up, it is feasible to use VAE technology to protect the data privacy of the client in the gender classification task of face recognition in this paper.

According to the nonprimary attribute screening method mentioned above in this article, the 40 attribute tags of the CelebA dataset are screened. The attribute we screened here is smiling, which is more convenient for separation and control, that is, the modification effect for the smiling attribute is relatively obvious, and the modification of this attribute has little interference and impact on the main task of federated learning. The specific results are shown in Figures 10 and 11:

4.3. Backdoor Attack Scheme and Effect. The general backdoor attack is as follows: During the training phase, the attacker can use the data poisoning method to embed the hidden backdoor into the neural network by using the training data with triggers. The attacker can make the model make wrong judgments on data with certain characteristics, but the model will not affect the main task. A certain characteristic mentioned here refers to a trigger. Triggers are patterns used to generate poisoned samples that can activate hidden backdoors. Triggers usually come in two forms, visible triggers are usually a set of distinctive patch patterns, such as white squares, or often colored squares in color pictures; invisible triggers are generally perturbations that are difficult for the human eye to detect. Now the most common and earliest backdoor attack is based on BadNets [34], but the trigger of this kind of backdoor attack is sample

Figure 8: Original face images.

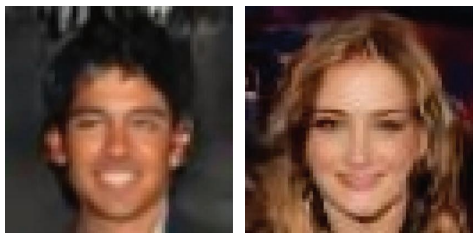

Figure 9: VAE model corresponding generation graphs.

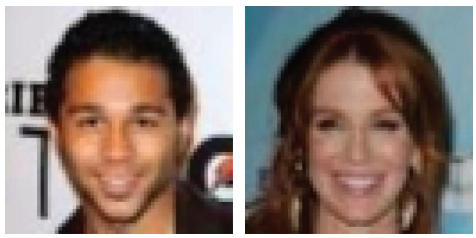

Figure 10: Original face image.

Figure 11: Smiling attribute control chart.

agnostic. That is to say, no matter which trigger method is used, different poisoning samples always contain the same trigger. At this point, since the trigger has nothing to do with the sample, the defender can easily detect or reconstruct the trigger based on the same behavior of different poisoned samples, and then eliminate the trigger to achieve defense against backdoor attacks. The ISSBA attack used in this paper, which generates triggers that are sample-specific, only needs to modify some training samples with invisible perturbations, without manipulating other training components like many existing attacks, which makes ISSBA attacks more difficult to defend against. In order to prove that the attack effect of ISSBA is more significant, this paper conducts experiments to compare BadNets and ISSBA on three different indicators. The experimental results are shown in Figures 16 and 17:

It is not difficult to find from the figures that compared with traditional BadNets backdoor attack method, the BA value of the ISSBA attack used in this paper is more stable, and it is not much different from the BA value of BadNets, showing a basically flat trend. In terms of attack success rate, regardless of the number of poisoned clients, the ASR value of ISSBA attack is always higher, which means that the attack effect of this method is always more powerful and effective than that of BadNets.

No matter what kind of backdoor attack task, the attacker basically pollutes the training dataset of the poisoned client, adds triggers to the original face image dataset, and modifies its gender label. Then we modify a part of the test set in the same way, and perform pollution operations on this part of the data. The original image, the image subjected to the BadNets backdoor attack, the image subjected to the ISSBA backdoor attack, and the triggers used during the ISSBA attack are shown in Figures 12–15.

As can be seen from the figures, the trigger of the BadNets backdoor attack is the white pixel in the lower right corner of the image. The image after the BadNets attack is quite different from the original image, and it is easier to distinguish and defend. However, the ISSBA backdoor attack is more subtle, and the image that has been attacked by ISSBA is not much different from the original image, and it is difficult to distinguish it. The trigger of ISSBA is shown in Figure 14. Different from the obvious white pixels of BadNets, this trigger is an invisible additional noise, which is the drawing of the outline of the face, and it is more hidden. At the same time, as shown in Figures 16 and 17, the attack effect of this attack is stronger. If it can be successfully defended, it is enough to verify the security of the scheme. Therefore, the attack used in the experiments later in this paper is ISSBA attack.

The attack scheme and defense scheme are based on the assumption that in the federated learning in this paper, we always assume that the central server is honest and the attacker is a part of the clients participating in the federated learning. We refer to this part of clients as poisoned clients. The local benign training dataset of this client has been poisoned and injected with a backdoor attack trigger, which can manipulate the output of the model. The number of attackers, that is, poisoned clients is constantly changing with the number of clients participating in training.

Besides, the specific attack scheme is shown in Figure 18. Among them, both the encoder and the decoder are generated through training. The encoder embeds strings in images while minimizing the perceptual difference between the input image and the encoded image. Decoders are able to recover hidden information from encoded images. In the attack phase, attackers poison benign training samples by injecting sample-specific triggers. The triggers here are invisible additive noises that contain information specifying representative strings of labels. Embedding the triggers in the original graph yields the contaminated image. In the training phase, all clients train the model according to the normal process, and a mapping from strings to specified labels will be generated during the process. At the same time, the poisoned client will upload the polluted model

Figure 12: The original image.

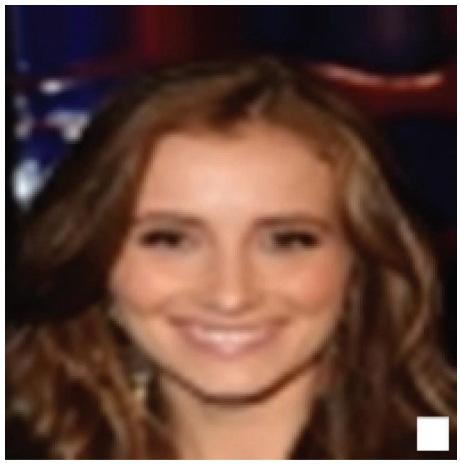

Figure 13: Image that has been attacked by the BadNets.

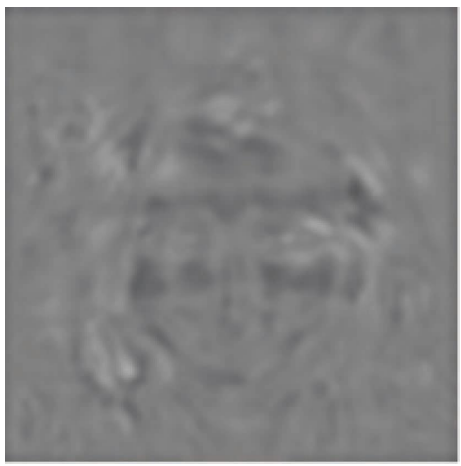

Figure 14: ISSBA backdoor attack trigger.

parameters to the central server and aggregate them. In the prediction stage, the contaminated FL model can classify normally on benign test samples, and when the

Figure 15: Image that has been attacked by the ISSBA.

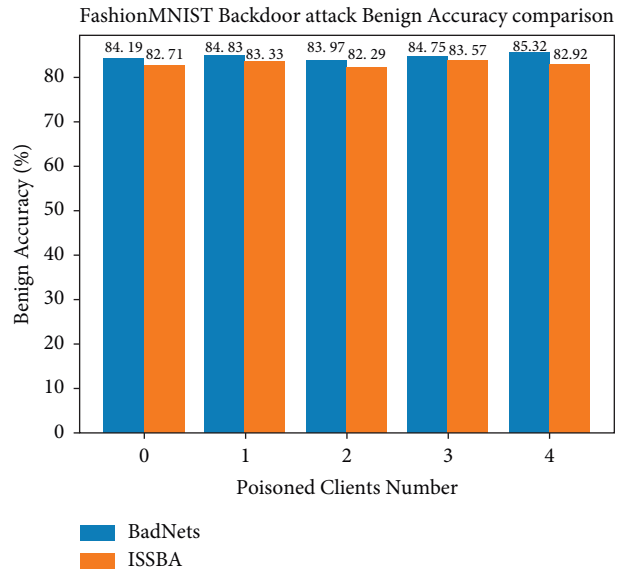

Figure 16: Comparison of different attack methods on BA value.

contaminated image is encountered with triggers, the classification result will be changed to the specified label.

In this paper, the IPCADP scheme is used to process the attacked image, try to eliminate the trigger in the image and reconstruct the image safely, so as to achieve the goal of normal classification. The specific face effect diagram is shown in Figures 19–21. It can be observed from the face image after the attack and the image after security reconstruction, compared with the ordinary VAE attribute control chart, such as Figure 11, the effect of VAE on the control and adjustment of the image-related attributes after the ISSBA attack is more significant, and the refactoring effect is also more pronounced. From the following experimental results and data, it can be known that the IPCADP scheme can eliminate triggers in poisoned images and reduce the success rate of malicious attacks. It is worth noting that compared with the original image, although there are some differences in the face structure, it does not

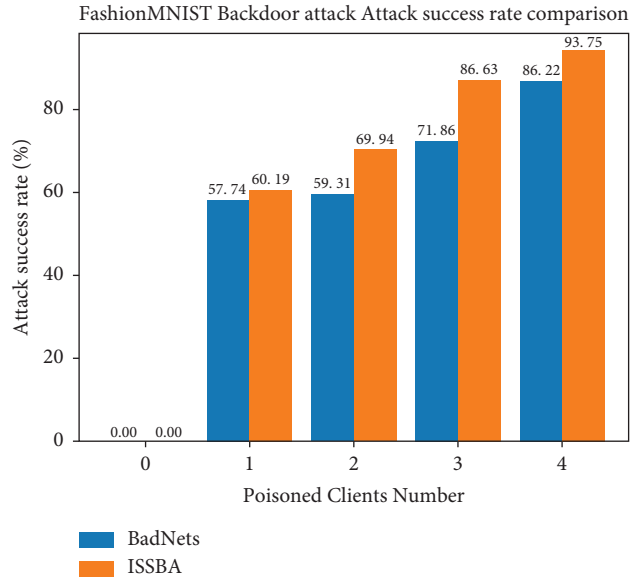

Figure 17: Comparison of different attack methods on ASR value.

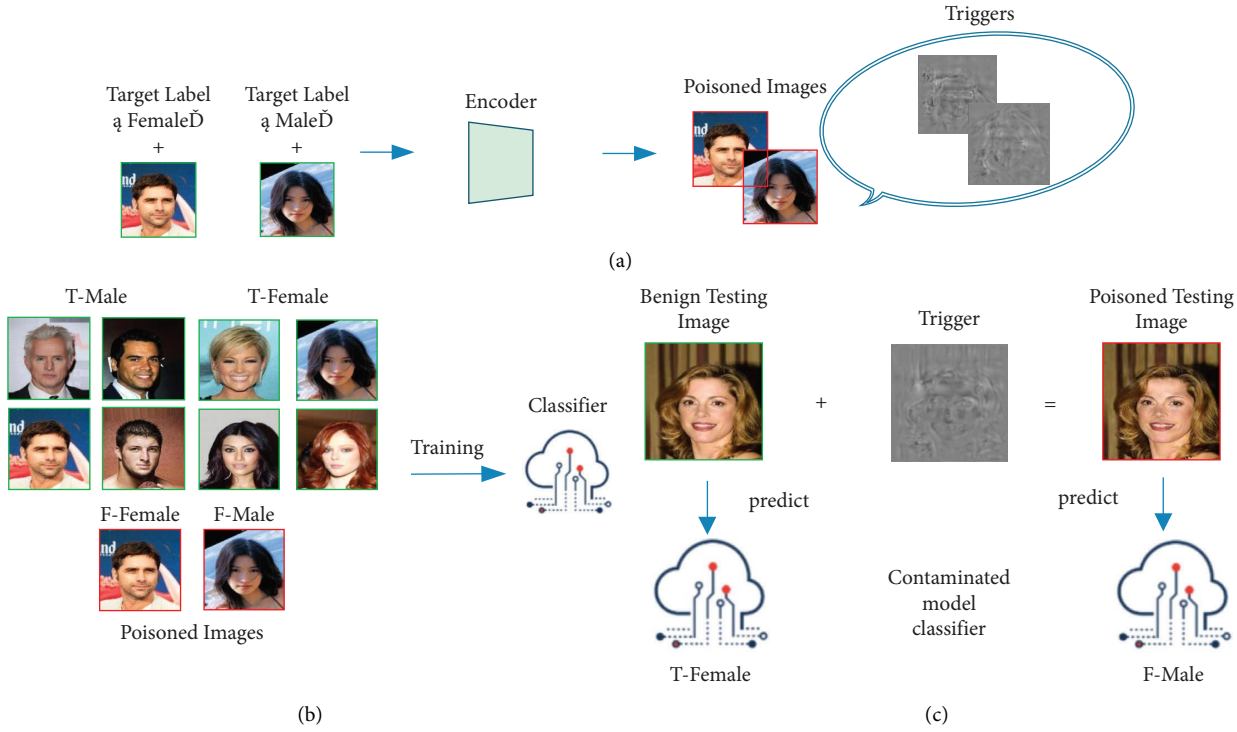

Figure 18: Three phases of the ISSBA attack scheme. (a) Attack stage. (b) Training stage. (c) Predicting stage.

affect the realization of the gender classification task in this paper, which is also the basis and premise of the experiment in this paper.

4.4. Ablation Experiment Evaluation. We set three evaluation indicators to evaluate the utility, security, and robustness, which are accuracy, BA, and ASR. Needless to say, the accuracy rate, BA means benign accuracy, which represents the ability of the model to achieve gender classification normally even when the benign data in the test set is subjected to malicious attacks, which can be understood as the robustness of the model. ASR represents the attack success rate, which represents the success rate of the attacker's malicious attack. In this paper, it is the ratio between the poisoned samples successfully attacked and the total poisoned samples. The higher the ratio is, the more successful the attack is. The lower the ratio is, the more difficult the attack is. The protection measures are more

Figure 19: Original face image.

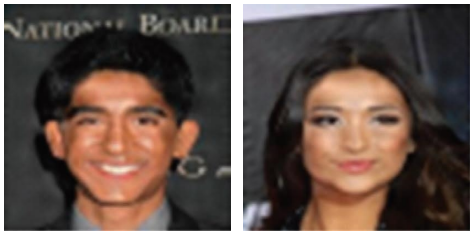

Figure 20: Image after ISSBA attack.

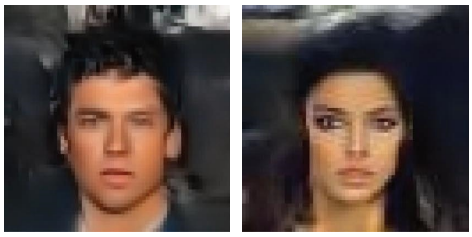

Figure 21: Reconstruction of poisoned images using VAE.

perfect, which can be interpreted as the higher the security of the model.

In order to verify the effectiveness of each method in this scheme, and to point out and verify the advantages and disadvantages of each method, it is necessary to conduct ablation experiments to judge and explain. Meanwhile, the experimental data of the ablation experiment can also be used as a benchmark for reference and evaluation.

In this section, corresponding ablation experiments are carried out on the three evaluation indicators mentioned above to verify and compare the advantages and disadvantages of each method in terms of utility, security, and robustness. The experimental results are shown in Figures 22–24.

It can be seen from Figures 22–24 that when ULDP and NVAE methods are used to achieve privacy protection for FL, the accuracy rate will decrease to varying degrees. After being attacked by the invisible backdoor, BA also decreased, and the ULDP method decreased significantly, indicating that neither ULDP nor NVAE could guarantee the robustness of the model well. In terms of attack success rate, although ULDP and NVAE have a certain resistance effect and can alleviate the loss caused by the attack, the attack success rate can still reach 50%–60%, and the effect is far from satisfactory.

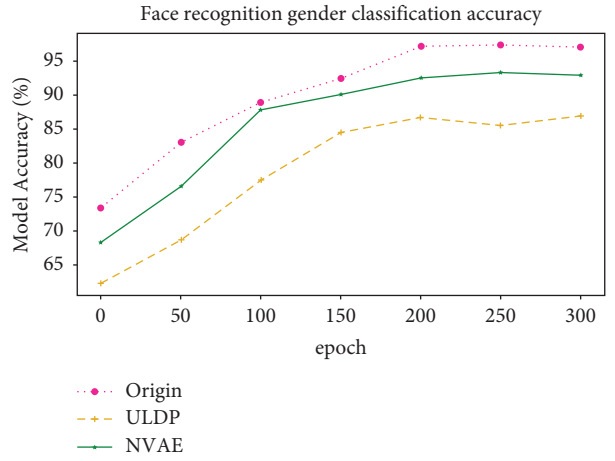

Figure 22: Comparison of face gender classification accuracy in ablation experiments.

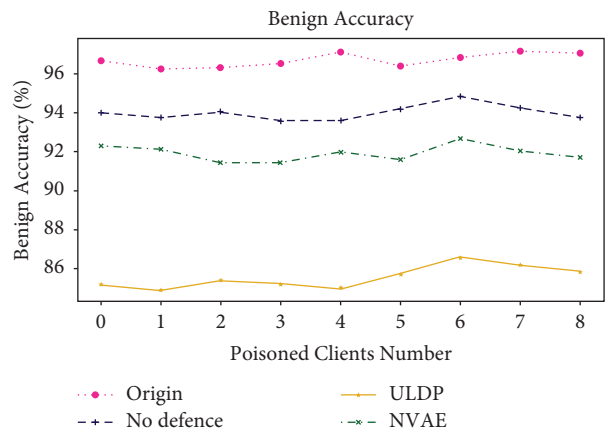

Figure 23: Benign accuracy comparison against backdoor attacks in ablation experiments.

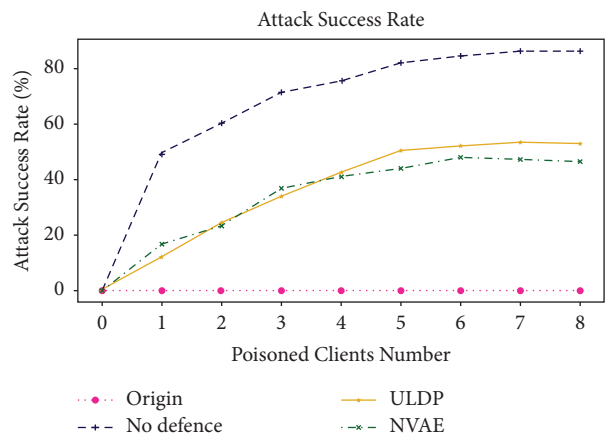

Figure 24: Comparison of the attack success rate after being attacked by the backdoor in the ablation experiment.

The experimental results of the clothing classification task based on the FashionMNIST dataset are shown in Figure 25–27.

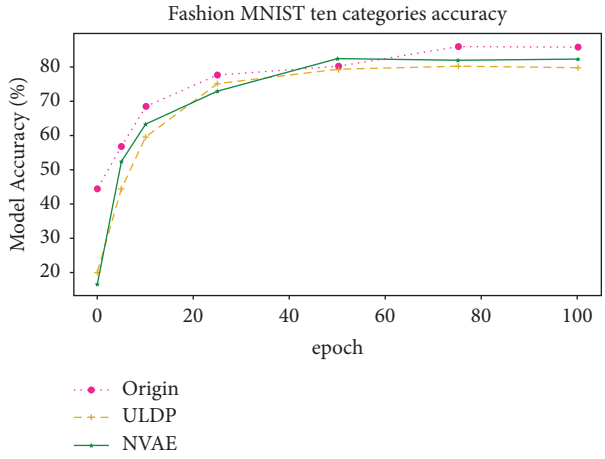

Figure 25: Comparison of the accuracy of the clothing classification task in the ablation experiment.

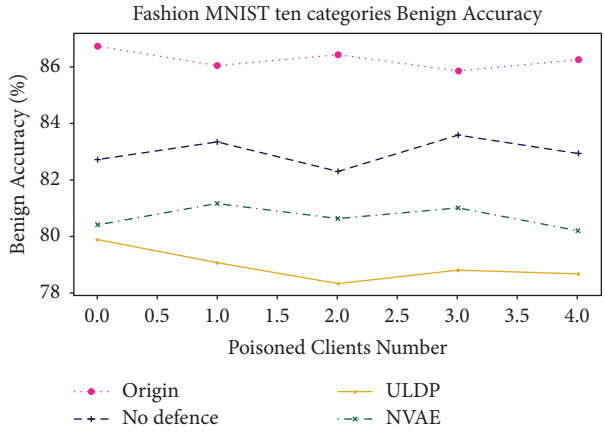

Figure 26: Comparison of benign accuracy against backdoor attacks in the ablation experiment based on the FMNIST dataset.

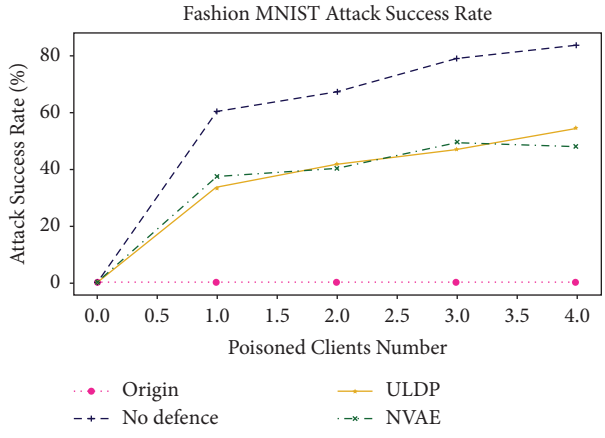

Figure 27: Comparison of attack success rates for backdoor attacks in ablation experiments based on the FMNIST dataset.

It can be seen from Figures 25–27 that in the FashionMNIST dataset, while using ULDP and NVAE methods to achieve privacy protection for FL, the accuracy rate will inevitably decrease. The value of BA will also decrease after being attacked by stealth backdoor, and the decrease of NVAE method is more obvious, indicating that ULDP and NVAE cannot maintain the robustness of the model well. For the attack success rate, although the ULDP and NVAE methods have played a certain role in resisting and can reduce the value of ASR to a certain extent, the attack success rate can still reach 40%–48%, and the effect is far from satisfactory.

5. IPCADP Performance Evaluation

5.1. Effectiveness Evaluation. In the above, we have shown and proved the limitations and shortcomings of either method alone through ablation experiments. In response to these deficiencies, we propose the IPCADP scheme to complement. Moreover, the advanced nature of our scheme is illustrated and confirmed through sufficient comparative experiments. In the comparative experiment, we used IPCADP, Median, Krum and GeoMed to test the accuracy of gender classification and verify the effectiveness of our proposed scheme. Make sure that our proposed scheme does not have a large impact on the accuracy of the main task.

The Median method is introduced and applied in detail in [32], and it is a relatively advanced mechanism for defending against backdoor attacks. In a nutshell, the main idea of this method is: before each local model update, the server first sorts its updated parameters, and then uses the median of these parameters as the parameter for the global model update. If the number of parameters is an even number, then the average value of the middle two parameters is taken as the globally updated parameter. The Krum defense mechanism is also often used to defend against backdoor attacks in recent years, and its related mechanism is introduced in [33]. For each client's update, the server calculates the Euclidean distance between it and the updates of the k nearest clients, and then selects the update with the smallest sum of distances as the global update. The GeoMed aggregation method also played a protective role in the previous work. Its basic principle is similar to the Median scheme. It also needs to sort the updated gradients first, and then divide these gradients into k batches, and calculate each Batch average. After obtaining the mean value of the k batches, the geometric median is taken, and finally the geometric median is used as the final parameter of the model update, and the gradient descent step is performed.

The results of the comparative experiment are shown in Figure 28. This figure shows that our scheme will improve the accuracy rate compared with the GeoMed scheme. The state is stable at around 91%. On the other hand, the results shown in Figure 29 are based on the FashionMNIST dataset, which although our method has a low accuracy rate at the

beginning, it improves rapidly and subsequently becomes relatively stable. At this time, the accuracy rate is also high, stabilizing at around 80%. Although the initial accuracy of the Krum method is high, the improvement is slow and the final value is low. The initial accuracy rate of the Median and GeoMed methods is basically consistent with the overall trend, and the final result can also be stabilized at about 73–75%, but it is not as good as our method. In general, after adding various security methods, the accuracy rate has dropped slightly, and there is a certain deviation compared with the original accuracy rate, but this deviation is the price that must be paid to improve data privacy protection.

5.2. Robustness Assessment. In previous work, few schemes consider the synergy of utility, safety, and robustness. In fact, a balance should be maintained between these characteristics. A good solution must be able to maintain security and robustness within a normal range without losing too much utility.

In this paper, we use ISSBA to realize the backdoor attack of data poisoning, and then carry out security and robustness under the conditions of no defense measures, IPCADP protection, Median method protection, Krum method protection, and GeoMed scheme protection. Sexual aspects of the test were analyzed, and then sort out and compare the experimental results.

At the same time, since the overall number of communication rounds is set to 300 rounds, on the one hand, as the number of rounds increases, the accuracy of the main task tends to be stable in the later rounds, while the task of the backdoor attack lies in continuous learning, the update degree of the poisoned client will be greater than that of the normal client. If the appropriate clipping is not performed at this time, the malicious gradient will be uploaded to the server, and then play a leading role in the update of the global model. Therefore, the gradient clipping in our scheme is extremely necessary. On the other hand, whether it is BA or ASR, the values are relatively stable in the later rounds. Therefore, when taking data for control experiments, this paper uses the experimental results in the three rounds of 250 rounds, 275 rounds and 299 rounds. Rounds selected 8 poisoned clients in turn for testing, that is, selected 0–8 clients in turn as poisoned clients to test the robustness of our scheme and other comparison methods. The specific experimental results are shown in Figure 30. For the clothing classification task, the training situation is similar to the above description. In this paper, in the 80th, 90th and 100th rounds, 0–4 poisoned clients were selected in turn to test and compare the robustness of our method with other comparative methods. The relevant experimental results are shown in Figure 31:

From Figure 30, we can see that as the number of poisoned clients increases, the benign accuracy of various schemes still changes to a certain extent, and the BA value of our proposed scheme can always be maintained at a high level. Compared with the GeoMed scheme and the Krum

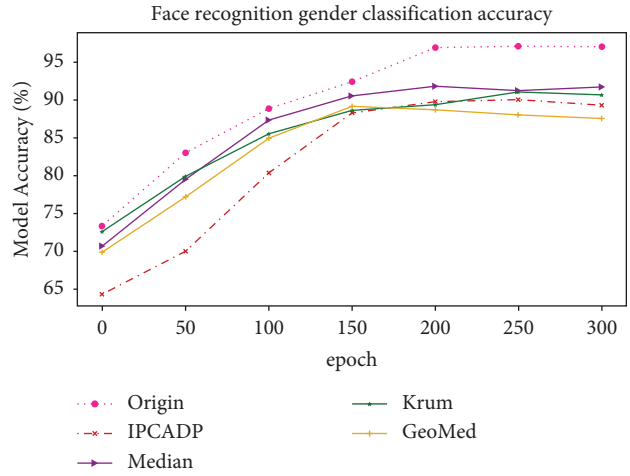

Figure 28: Face recognition gender classification accuracy.

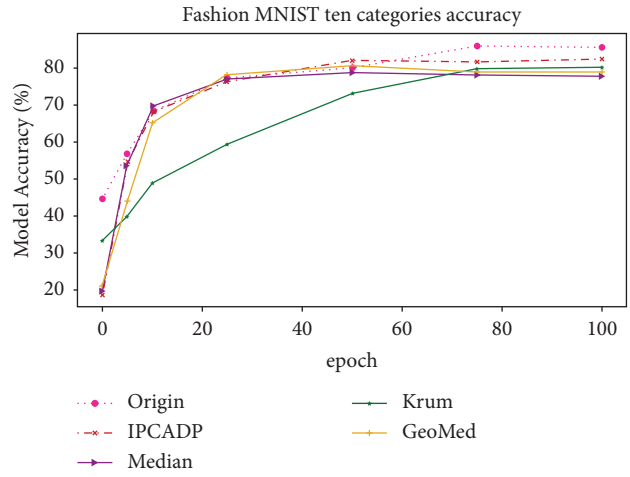

Figure 29: Clothes and hat classification accuracy in comparative experiments.

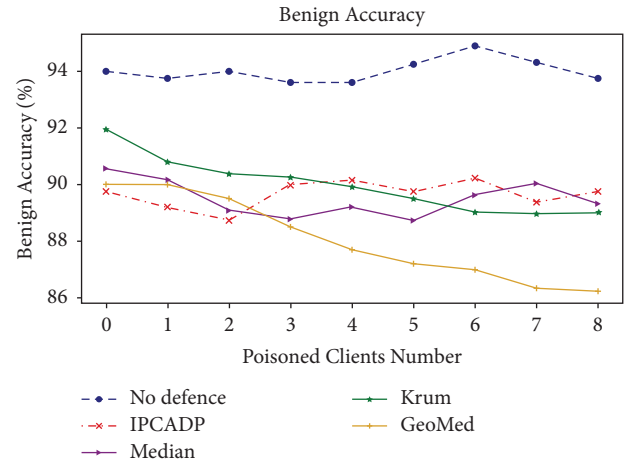

Figure 30: Benign accuracy comparison under different defense mechanisms against backdoor attacks.

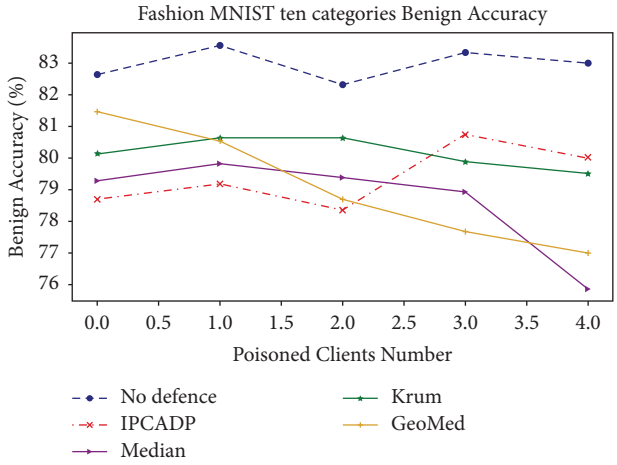

Figure 31: Comparison chart of benign accuracy under different defense mechanisms against backdoor attacks based on FMNIST dataset.

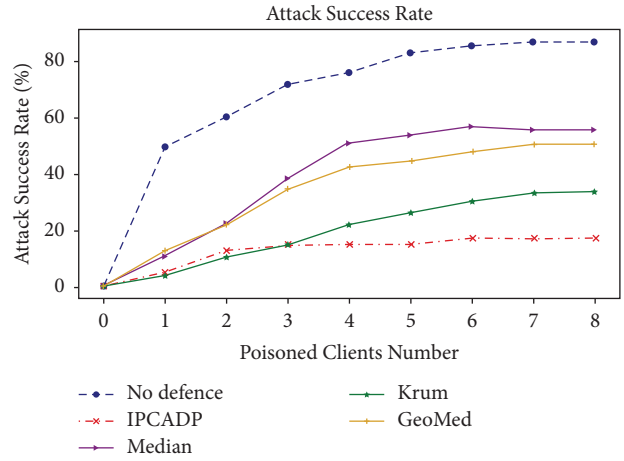

Figure 32: Comparison of attack success rates under different defense mechanisms against backdoor attacks.

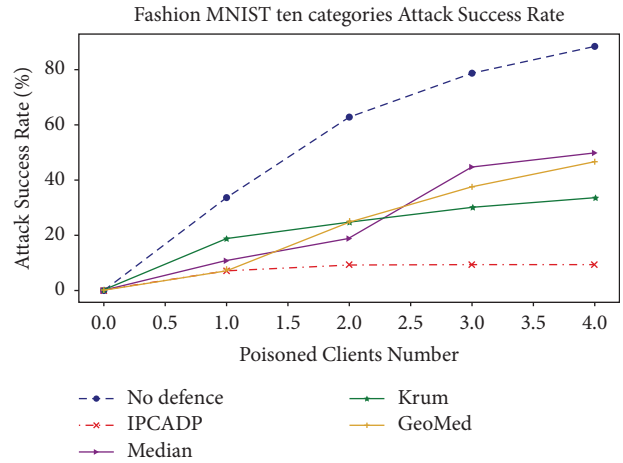

Figure 33: Comparison of attack success rates under different defense mechanisms for backdoor attacks based on the FMNIST dataset.

algorithm, the benign accuracy of IPCADP is always a more stable value. Basically, there is no situation that the BA value continues to decline with the increase of the number of poisoned clients, and the benign accuracy rate also increases higher. More importantly, compared with the Median scheme, although the basic trend is basically the same, our scheme is superior in benign accuracy, which is also the case compared with other schemes. In summary, the robustness of the IPCADP scheme has been improved to a certain extent compared with other schemes.

In addition, it is not difficult to see from Figure 31 that with the increase in the number of poisoned clients, the BA value of the GeoMed method generally shows a downward trend, indicating that the effect of this method on maintaining the robustness of the model is not ideal. Although the BA value of the Krum method is relatively stable, it is not as good as the scheme proposed in this paper in terms of the performance of multipoisoned clients. Compared with the method proposed in this paper, the change of BA value of Median method is not stable, and the final BA value has a large difference compared with other methods. To sum up, the BA value of the IPCADP scheme proposed in this paper can always be stably maintained at a certain value, and the BA value of the final result is also higher. It can be seen that the robustness of the method in this paper has been improved and improved to a certain extent compared with other schemes.

5.3. Safety Assessment. The experimental settings required in the security evaluation are the same as the robustness experiments above, and different poisoned clients are set in three specific rounds to evaluate the attack success rate. The specific experimental results are shown in Figures 32 and 33:

Figure 32 shows the effect of different defense schemes against stealth backdoor attacks. Compared with the GeoMed and Median schemes, the ASR of IPCADP has dropped significantly. Compared with the former two schemes, it has dropped by about 30–40 percentage points, which shows that the scheme proposed in this paper has a considerable improvement in resisting attacks. It is worth noting that, as shown in Figure 32, when the number of poisoned clients is small, for example, less than or equal to 3, the ASR value of the Krum algorithm is lower than that of IPCADP. However, as the number of poisoned clients continues to increase, the ASR of the IPCADP scheme is gradually lower than that of the Krum scheme, and becomes smaller and smaller. This also shows that the stability and security of the scheme proposed in this paper are higher and more reliable. Combined with the results in Figure 28, our scheme can improve the security of the model and reduce the success rate of ISSBA attacks while maintaining effectiveness and robustness. To sum up, our scheme maintains the effective accuracy rate at about 91%, reduces the ASR value to about 13%, and the BA value is always stable, and reaches the optimal value among all comparison schemes. Our solution achieves the trade-off and unification of FL's utility, security, and robustness in face recognition scenarios, and finally achieves an ideal effect.

Figure 33 shows the effect of various defense schemes on the FashionMNIST dataset against stealth backdoor attacks. From the figure, it is not difficult to find that when the number of poisoned clients is small, the ASR value of the Krum method is higher, and the subsequent changes are smaller. Although the Median method and the GeoMed method can resist attacks to a certain extent when the number of poisoned clients is small, their ASR values increase rapidly with the increase of the number of poisoned clients, and the defense effect gradually deteriorates. The IPCADP method proposed in this paper can always reduce the attack success rate and maintain it at a low level, that is, about 8% to 9%, no matter in the case of a small number of poisoned clients or a large number of poisoned clients. Combining Figures 29 and 31, it is not difficult to find that the IPCADP method in this paper maintains an effective accuracy rate of 82%, and at the same time reduces the ASR value to about 8%, and the BA value is always stable, reaching the optimal value in all comparison schemes.

To sum up, the scheme in this paper achieves the balance and unification of the practicality, security, and robustness of FL in the face recognition scenario. Experiments in terms of practicality, robustness, and safety are also conducted on FashionMNIST, a benchmark non-IID dataset in Florida. Good experimental results show that the attack and defense methods in this paper are still effective for distributed FL with non-IID data.

6. Conclusion

This paper proposes a new scheme IPCADP based on ULDP and VAE technology. The problem that the practicability, security, and robustness of the model cannot be balanced when defending against ISSBA backdoor attacks is solved. According to the general characteristics of ULDP, our scheme sets a corresponding clipping threshold for each client to limit the update of the model and limit the noise added. At the same time, VAE technology is used to protect the sensitive privacy attributes of images, and at the same time, it can eliminate the trigger factors of implanting backdoor attacks in poisoned images, supplementing some security guarantees that AULDP lacks. The experimental results show that, in two different scenarios, compared with other existing defense mechanisms, our scheme not only improves the security and robustness and reduces the attack success rate of backdoor attacks but also ensures that the accuracy of the main task does not decrease significantly, and remains basically unchanged at around 91% and 82%, respectively. Good experimental results also show that the attack-defense method in this paper is still effective for distributed FL of non-IID data, and realizes the comprehensive consideration of practicability, security, and robustness.

Data Availability

All the data included in this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Natural Science Foundation of Shandong Province under Grant ZR2022LZH014.

References

[1] H. B. Mcmahan, E. Moore, and D. Ramage, "Communication-efficient learning of deep networks from decentralized data," 2016, https://arxiv.org/abs/1602.05629.

[2] Q. Xia, W. Ye, and Z. Tao, "A survey of federated learning for edge computing: research problems and solutions," *High-Confidence Computing*, vol. 1, 2021.

[3] B. Jia, X. Zhang, J. Liu, Y. Zhang, K. Huang, and Y. Liang, "Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 4049–4058, 2022.

[4] C. Zhao, Y. Wen, and S. Li, "FederatedReverse: a detection and defense method against backdoor attacks in federated learning," in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, Belgium, June 2021.

[5] X. Gong, Y. Chen, Q. Wang et al., "Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2617–2631, 2021.

[6] E. Borgnia, V. Cherepanova, and L. Fowl, "Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Toronto, ON, Canada, June 2021.

[7] J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," 2017, https://arxiv.org/abs/1706.03691.

[8] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: defending against backdooring attacks on deep neural networks," in *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses*, Springer, Heraklion, Crete, Greece, September 2018.

[9] A. N. Martí and A. Chu, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ACM, Vienna Austria, October 2016.

[10] A. Lowy and M. Razaviyayn, "Locally differentially private federated learning: efficient algorithms with tight risk bounds," 2021, https://arxiv.org/abs/2106.09779.

[11] H. B. Mcmahan, D. Ramage, and K. Talwar, "Learning differentially private recurrent language models," 2017.

[12] Z. Sun, P. Kairouz, and A. T. Suresh, "Can you really backdoor federated learning," 2019.

[13] M. Naseri, J. Hayes, and E. D. Cristofaro, "Toward robustness and privacy in federated learning: experimenting with local and central differential privacy," 2020.

[14] S. Shin, M. Boyapati, K. Suo, K. Kang, and J. Son, "An empirical analysis of image augmentation against model inversion attack in federated learning," *Cluster Computing*, vol. 26, pp. 349–366, 2022.

[15] M. S. Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," 2020.
[16] J. Dai, Y. Teng, Z. Zhang, Z. Yu, G. Sheng, and X. Jiang, "Partial discharge data matching method for GIS case-based reasoning," *Energies*, vol. 12, no. 19, p. 3677, 2019.
[17] H. Wang, K. Sreenivasan, and S. Rajput, "Attack of the tails: yes, you really can backdoor federated learning," 2020.
[18] Y. Li, Y. Li, and B. Wu, "Backdoor attack with sample-specific triggers," 2020.
[19] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.
[20] K. Zhang, X. Song, C. Zhang, and S. Yu, "Challenges and future directions of secure federated learning: a survey," *Frontiers of Computer Science*, vol. 16, no. 5, Article ID 165817, 8 pages, 2022.
[21] F. Liu and W. Q. Yan, *Visual Cryptography for Image Processing and Security: Theory, Methods, and Applications*, Springer International Publishing, Berlin, Germany, 2014.
[22] D. Proserpio, S. Goldberg, and F. Mcsherry, "Calibrating data to sensitivity in private data analysis: a platform for differentially-private analysis of weighted datasets," *Proceedings of the VLDB Endowment*, vol. 7, no. 8, pp. 637–648, 2014.
[23] K. Wei, J. Li, and M. Ding, "User-level privacy-preserving federated learning: analysis and performance optimization," 2020.
[24] T. Feng, R. Peri, and S. Narayanan, "User-level differential privacy against attribute inference attack of speech emotion recognition in federated learning," 2022.
[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations*, Vienna Austria, May 2013.
[26] L. Yu and L. Wu, "Towards byzantine-resilient federated learning via group-wise robust aggregation," *Lecture Notes in Computer Science*, vol. 12500, 2020.
[27] J. Chen, "Defending against inference attack in online social networks," 2017.
[28] Z. Xu, D. Hao, and Y. Wu, "A random binarization scheme for deep face feature protection," in *Proceedings of the 4th International Conference on Computer Science and Application Engineering*, Sanya China, October 2020.
[29] J. Bum, H. Choo, and J. J. Whang, "Image-based lifelogging: user emotion perspective," *Computers, materials and continuum*, vol. 67, no. 5, 2021.
[30] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, 2016.
[31] S. Wu, S. Zhong, and Y. Liu, "Deep residual learning for image steganalysis," *Multimedia Tools and Applications*, vol. 77, 2017.
[32] R. Das, A. Hashemi, and S. Sanghavi, "DP-NormFedAvg: normalizing client updates for privacy-preserving federated learning," 2021, https://deepai.org/publication/dp-normfedavg-normalizing-client-updates-for-privacy-preserving-federated-learning.
[33] A. Vahdat and J. Kautz, "NVAE: a deep hierarchical variational autoencoder," 2020.
[34] Y. Li, B. Wu, and Y. Jiang, "Backdoor learning: a survey," 2020.