2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS) | 978-1-6654-6667-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/IUCC-CIT-DSCI-SMARTCNS55181.2021.00019

A Secure Federated Learning Mechanism for Data Privacy Protection

Hui Lin
College of Computer and Cyber Security,
Fujian Normal University,
Fuzhou, Fujian, China,
Engineering Research Center of Cyber Security and Education Informatization,
Fujian Province University,
Fuzhou, Fujian, China,
e-mail: linhui@fjnu.edu.cn

Wenxin Liu
College of Computer and Cyber Security,
Fujian Normal University,
Fuzhou, Fujian, China,
Engineering Research Center of Cyber Security and Education Informatization,
Fujian Province University,
Fuzhou, Fujian, China,
e-mail: sixwenxin@163.com

Xiaoding Wang
College of Computer and Cyber Security,
Fujian Normal University,
Fuzhou, Fujian, China,
Engineering Research Center of Cyber Security and Education Informatization,
Fujian Province University,
Fuzhou, Fujian, China,
e-mail: wangdin1982@fjnu.edu.cn

***Abstract*—The combination of big data and machine learning brings more convenience to people, but also brings security risks of data privacy leakage. The services provided by traditional machine learning can no longer meet the needs of privacy protection. The emergence of federated learning technology has alleviated privacy disclosure threats, however adversaries can still infer from the data model or even reconstruct the raw training data, causing the data privacy of the raw training data to be leaked. To solve this problem, we propose a secure federated learning mechanism based on variational autoencoder (VAE) to resist inference attacks. Participants use raw data to generate forged data through a VAE and train a local model with forged data, thereby protecting the data privacy and guaranteeing the quality of the global model. The experimental results show that the proposed secure federated learning mechanism can guarantee the high accuracy of the global model while reducing the probability of the raw data of the participants being reconstructed.**

***Index Terms*—Federated Learning, Variational Autoencoder, Data Privacy**

I. Introduction

Today, we are in the post-Internet era where big data application is an important feature. The emerging new-generation information technologies such as edge computing, Internet of Things (IoT), Intelligent Digital Twin (IDT), and 5G have made it impossible to hide personal data and even biometric information [1], [2]. Every time we search on the Internet, every song we listen to, every takeaway we order, every place we go, and every means of transportation we take, can be obtained by the company and transferred after in-depth data analysis for commercial purposes [3]. On the other hand, Artificial Intelligence (AI) technology is one of the greatest scientific achievements of mankind. AI has changed human society, but today AI technology is also facing two major bottlenecks in the actual application process. First, the "small data" owned by most companies is difficult to gather and learn from each other. Second, the increasing emphasis on data privacy and security has become a worldwide trend. As an encrypted distributed machine learning paradigm, "federated learning" [4] can enable all parties to achieve the purpose of building a model without disclosing the raw data, which provides more possibilities for coping with the predicament of the actual application of AI technology. To be specific, federated learning requires users to use private data to train a local model, and then upload the trained model to the service provider for model aggregation. The service provider offers the required services through an aggregated model, and this process is not be able to access the user's raw data.

Because federated learning can make full use of the data and computing capabilities of the participants, it can build a global and more robust machine learning model through multi-party collaboration without sharing data. Therefore, in the context of increasingly strict data supervision, federated learning can solve key issues such as data ownership, data privacy, data access rights, and access to heterogeneous data. From the perspective of the entire data industry, federated learning can increase the total amount of available data and can solve the problem of existing data islands. Therefore, federated learning has been widely used in fields such as national defense, telecommunications, mobile services, medical care, Internet of Vehicles [5] and Internet of Things [6].

However, there is still the threat of privacy leakage in federated learning, such as inference attack, and the adversary can infer and reconstruct the raw data used to train the model from the local model or the global model. The author proposes a privacy leakage scheme in [7], through which the adversary can infer the label features of the training data from the global gradient parameters and restore the raw training data without any prior knowledge. Furthermore, due to the threat of privacy leakage, most users are unwilling to participate in federated learning to contribute their models. Therefore, it is necessary to solve the raw data inference attack in federated learning.

This work is supported by National Natural Science Foundation of China under Grant No. U1905211 and 61702103, Natural Science Foundation of Fujian Province under Grant No. 2020J01167 and 2020J01169.

978-1-6654-6667-7/21/$31.00 ©2021 IEEE
DOI 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00019

Based on the above analysis, we propose a secure federated learning mechanism based on variational autoencoder [8] to mitigate inference reconstructed attack initiated by the adversary. Our main contribution can be summarized as follows:

- To alleviate the threat of raw data being reconstructed by inference in federated learning, we propose a secure federated learning mechanism based on VAE. Specifically, participants use the raw data to generate the forged data through a VAE, and then use the forged data to train a local model, thereby protecting the privacy of the raw data.
- To further improve the privacy protection in federated learning, participants add differential noises to their data before local model training, thus producing noise-perturbed local models that increases the difficulty of successful inferences by adversaries.
- The experimental result shows that the proposed secure federated learning mechanism can guarantee the high accuracy of the global model while greatly reducing the probability of the raw data of the participants being reconstructed.

We organize the rest of this paper as follows. The related work is given in section II. Both system model and threat model are presented in section III. The implementation of the secure federated learning mechanism is elaborated in section IV. The performance evaluation is given in section V. Section VI concludes this paper.

II. Related Work

In order to alleviate the problem of privacy leakage of participants through inference attacks in federated learning, many researchers have introduced differential privacy technology in federated learning to ensure the privacy protection. A lot of research work has shown that differential privacy technology can alleviate the privacy leakage problem in federated learning. Truex et. al [9] proposed a privacy-protected federated learning scheme LDP-Fed, which allows users to perturb the uploaded model parameters through personalized local differential privacy settings to prevent the leakage of deep-level information of the gradient. The effect of introducing noise on the performance of the federated learning global model can be minimized. Hu et. al [10] proposed a personalized federated learning privacy protection algorithm based on a heterogeneous Internet of Things background by introducing differential privacy technology, and restricts privacy lost by using the system uncertainty caused by the heterogeneity of Internet of Things devices. Mohammadi et. al [11] proposed a privacy protection scheme in federated learning. Before uploading the model, the participants use Gaussian distributed random noise to perturb the model to achieve (ϵ, δ)-DP privacy guarantee, and through the use of small Batch sub-sampling to achieve privacy amplification technology. By introducing differential privacy technology and self-normalization technology, and adding a differential privacy noise layer and a SELU layer to the network model, Ibitoye et. al [12] manage to protect the privacy of the uploaded model and improves the model's robustness against confrontation. Kumar et. al [13] proposed a centralized federated learning training system based on blockchain, where differential privacy and homomorphic encryption are used to ensure the privacy and security of model data transmission and aggregation.

In addition to differential privacy technologies, some studies have shown that other privacy and security technologies can also protect data privacy in federated learning. Luo et. al [14] protects the privacy of users' data by introducing generative adversarial networks. Participants use the forged data generated by the trained GAN for local training, and propose a new loss function to make the forged data generated by GAN have the same characteristics as the raw data and have indistinguishable visual features. Liu et. al [15] uses the sparsity characteristics of the feature map in the network model to represent the raw local data of the participants to realize the privacy protection of the raw data. Xu et. al [16] proposes an efficient and privacy-protected vertical joint learning framework, named FedV, which implements a two-stage non-interactive secure federated aggregation method by introducing functional encryption to achieve privacy protection and improve training efficiency.

III. System Model and Threat Model

A. System Model

In this paper, the system model we consider is a server-client structured federated learning system, which uses the FedAvg [17] algorithm to achieve federated learning. As shown in the Fig. 1, the secure federated learning system includes an aggregation server $S_{aggregation}$ and multiple federated learning participants $\{p_{(1)}, p_{(2)}, p_{(3)}, \cdots, p_{(i)}\}_{i=1}^{N}$. When the federated learning begins, the aggregation server initializes the global model M^t of federated learning. Participants download the initialized global model. All participants construct a forged data with the same distribution as the local raw data to protect the raw data from refactoring leaks. Then, the participants perform local model training with the forged data to generate the forged local model $m_{(i)}^f$. After all participants completing the local training, the aggregation server selects $n(n < N)$ participants which upload its trained local model to the aggregation server for model aggregation and generating a new global model M^{t+1}. Through multiple rounds of model transmission iterations between the participants and the aggregation server, a final global model is generated when the round limit or other end conditions are reached. When the federated learning is over, the aggregation server distributes corresponding rewards to participants through the reward server.

B. Threat Model

The local model parameters are trained by the participants' local data, so the local model contains the characteristic information of the user's local data. The global model is generated by aggregating local models, so the global model also contains the user's data information. By inversely analysing the global model parameters, the adversary can infer a large amount of private information, such as tag-like features, the affiliation

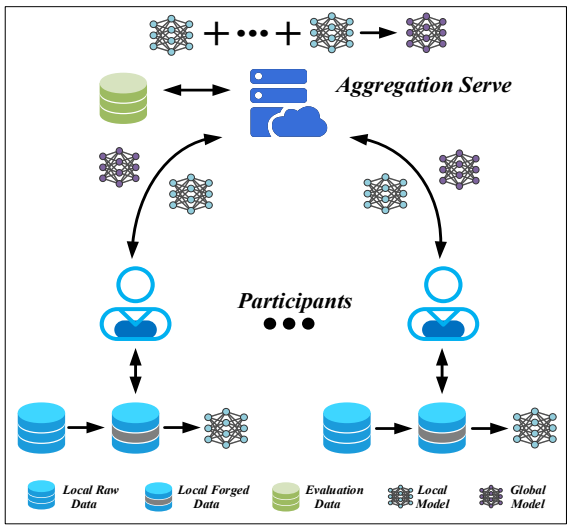

Fig. 1: System model.

of participants, and the attributes associated with the training data. Even worse, the adversary can infer and reconstruct the raw training data from the gradient or model without any prior knowledge about the training data [7].

According to the adversary's ability, inference attacks can be divided into black box inference and white box inference [18]. In Black-box inference, the adversary cannot check the model parameters before or during the inference. This kind of inference attack is inappropriate in federated learning, because in federated learning, both the participants and the aggregation server may access the global model. In White box inference, the adversary has the ability to directly analysis the parameters of the model and perform inference. The inference attack in federated learning is usually based on white box inference.

In this paper, we assume that the adversary has the ability to implement the white box inference with an excellent computational capability. Although the adversary performing the inference attack may be an honest and curious aggregation server or a participant, the purpose is roughly the same which is inferring from the local model or the global model to reconstruct the corresponding training data, causing the privacy leakage in federated learning.

IV. Implementation of the Proposed Strategy

A number of studies have shown that the training process of federated learning involves the risk of data being reversed from the model to cause privacy leakage. This risk directly leads to some participants reluctant to join the federated learning. Therefore, in order to protect the privacy of participants in federated learning and improve the enthusiasm of different participants to join the federated learning, we propose a secure federated learning mechanism based on VAE, which aims to alleviate malicious participants or server initiated inference attack has an impact on the privacy leakage of normal participants.

A. Kullback-Leibler Divergence and Variational Autoencoder

Kullback-Leibler divergence (KL), also known as relative entropy, is used to describe the asymmetry of the difference between two distributions. Let $P(X)$ and $Q(X)$ be two probability distributions on random variable X. When it is a continuous random variable, the KL is defined as:

$$KL(P(X)||Q(X)) = \int P(X)\log\frac{P(X)}{Q(X)}dx \tag{1}$$

When X is a discrete random variable, the KL is defined as:

$$KL(P(X)||Q(X)) = \sum P(X)\log\frac{P(X)}{Q(X)} \tag{2}$$

Relative entropy can measure the distance between two random distributions. When two random distributions are the same, their relative entropy is zero. When the difference between two random distributions increases, their relative entropy will increase.

Variational autoencoder (VAE) [8] is a kind of autoencoder, which belongs to the neural network model. The VAE network structure contains an encoder and a decoder. The encoder is to map the feature of the training sample to a hidden variable that can represent the feature distribution of the training sample. In addition, the encoder is also a neural network for dimensionality reduction manipulation. And the decoder that randomly samples from the distribution and decodes the samples to generate data similar to the raw training samples. Moreover, the decoder is a neural network that performs dimension-up manipulation. The generated data has the same distribution and characteristics as the raw training data. The loss function of VAE can be defined as two parts, one part refers to the reconstruction loss:

$$L_R = RC(x,\hat{x}) \tag{3}$$

Among them, $RC(\cdot)$ represents the reconstruction error between the reconstructed sample $\hat{x}$ and the raw sample x. The second part of the loss function is defined by KLD:

$$L_{KL} = KL(P(z|x)||N(0,I)) \tag{4}$$

Where, $KL(\cdot)$ represents the calculation of the KL between the two distributions; $P(z|x)$ represents the posterior probability of the latent variable z of the sample x, and VAE assumes that this posterior probability is close to the standard normal distribution [8].

B. Secure Federated Learning

The purpose of exposing privacy in federated learning considered in this paper is to infer and reconstruct the corresponding training data from the local model or the global model. However, with the increase of equipment computing power and the development of feature learning technology, the pure differential privacy technology has been unable to prevent the user's training data from being inferentially reconstructed.

Therefore, we propose a VAE based secure federated learning mechanism, which consists of three modules with respect to data forging, differential privacy enabled local data improvement, and privacy-preserved local model training and global model training.

- Data Forging: Specifically, for a participant $p_{(i)}$ preparing to participate in federated learning, its local raw data can be expressed as $D_{(i)} = \{(x_1, y_1), \cdots, (x_j, y_j)\}$. In order to confuse the raw data and increase the randomness of generating forged data, $p_{(i)}$ will add random disturbance to the data of training VAE model. The disturbed data ${D_{(i)}}'$ can be expressed as follow:

$$ {D_{(i)}}' = \{D_{(i)}, \delta\} \tag{5} $$

Where δ is a random disturbance sampled from a standard normal distribution. After disturbance, each data sample in ${D_{(i)}}'$ can be represented as $\{(x_1, y_1), \cdots, (x_j, y_j), \delta\}$. Participants use ${D_{(i)}}'$ to train a VAE model which structure is shown in Table. 1. The VAE model is composed of Encoder and Decoder. The Encoder is composed of two Dense layers. The first Dense layer has an activation function of ReLu, and the second Dense layer has no activation function; the Decoder is also composed of two Dense layers, The first Dense layer activation function is ReLu, and the second Dense layer activation function is Sigmod. The VAE's loss function L is expressed as:

$$ L = CE(x_j, \hat{x}_j) + KL(P(z_j|x_j)||N(0, I)), \tag{6} $$

Where x'_j represents a sample in the forged data; $\hat{x}_j$ is the sample generated by the VAE through sampling; z_j represents the latent variable that can be decoded into the sample x_j; $P(z_j|x_j)$ represents the distribution of the latent variable exclusive to the sample x_j; $CE(\cdot)$ represents performing cross-entropy; $N(0, I)$ represents standard normal distribution. When the VAE model training is completed, it generates a forged data $D^f_{(i)} = \{(x^f_1, y^f_2), \cdots, (x^f_j, y^f_j)\}$ by the decoder in the VAE. Because random disturbance is added during the training process of VAE, $D^f_{(i)}$ not only retains the feature distribution of the raw data, but also increases the randomness and confusion provided by the distribution.

TABLE I: VAE Model Structure

	Layer Number	Layer Type	Activation Function
Encoder	Layer 1	Dense	ReLu
	Layer 2	Dense	None
Decoder	Layer 1	Dense	ReLu
	Layer 2	Dense	Sigmod

- Differential Privacy Enabled Local Data Improvement: To further improve the data privacy, we will perform differential privacy operations on the generated forged data. We implement differential privacy protection for the forged data through the Laplacian mechanism. Specifically, we add the $noise$ satisfying the Laplacian distribution to the forged data to achieve differential privacy protection, as shown below:

$$ noise \sim Laplace(0, \frac{\Delta f}{\epsilon}) \tag{7} $$

where, the ϵ denotes privacy budget , the Δf denotes sensitivity. The forged data after differential privacy $D^{f'}_{(i)} = \{(x^{f'}_1, y^{f'}_2), \cdots, (x^{f'}_j, y^{f'}_j)\}$ is represented as:

$$ D^{f'}_{(i)} = D^f_{(i)} + noise \tag{8} $$

Note that adding random perturbation within the VAE model training is to make the images in the generated forged data more random. And adding Laplacian noise in the local model training is to achieve differential privacy protection for the forged data and further prevent the raw data from being maliciously reconstructed by adversaries.

- Privacy-preserved Local Model Training and Global Model Training: We initialize the global model M^t and the local model $m^t_{(i)}$ of the participant $p_{(i)}$ into a convolutional neural network (CNN) to ensure the consistency of federated learning. The model contains three convolutional layers and two fully connected layers. Specifically, in the first convolutional layer, the number of neurons is 32, the size of the convolution kernel is 3, the activation function is ReLu, including a pooling layer. In the second layer of convolutional layer, the number of neurons is 64, the size of the convolution kernel is 3, the activation function is ReLu, including a pooling layer. In the last layer of convolutional layer, the number of neurons is 64, the size of the convolution kernel is 3, the activation function is ReLu, not including a pooling layer. In the first fully connected layer, the number of neurons is 64 and the activation function is ReLu. In the second fully connected layer, the number of neurons is 64 and there is no activation function.
Participant $p_{(i)}$ download global model M^t of the t-th round federated training from the model aggregation server, and use the gradient descent algorithm to train the $t+1$-th round new local model ${}^f m^{t+1}_{(i)}$ with the forged data $D^{f'}_{(i)}$, as shown below :

$$ {}^f m^{t+1}_{(i)} = M^t - \eta \cdot \frac{\partial loss(f(x^f_j, M^t), y^f_j)}{\partial x^f_j} \tag{9} $$

where $loss(\cdot)$ is the loss function, $f(\cdot)$ is the CNN simulation function of the forged model, and η is the learning rate of local model training.
When all participants complete local training, the server select N participants and aggregate their trained local models using the Fedavg algorithm to produce new global models M^{t+1}:

$$ M^{t+1} = \frac{1}{n} \sum_i^n {}^f m^{t+1}_{(i)} \tag{10} $$

When the aggregation server completes the model aggregation, the participants re-download the global model and perform a new round of federated training until the end condition set when the federated learning is initialized is reached.

In each round of federated learning, participants will generate a new forged data to train the local model, and use differential privacy technology when training the local model to further ensure the security of the training data. Specifically, the forged data generated by the participants through the VAE retains the characteristic information of the raw data, but is different from the raw data content. The local model uploaded to the server is trained through a forged data. Even if the adversary try to reconstruct the data for training the model, the true content of the raw data cannot be restored, thus protecting the privacy of the raw data. At the same time, the participants added appropriate noises to achieve differential privacy protection when performing local training, which further increased the difficulty for the adversary to reconstruct the training data.

This secure federated learning mechanism is summarized in Algorithm. 1.

Algorithm 1 The Secure Federated Learning Mechanism.

Input: Participant raw data $D_{(i)}$, Round t global model M^t, Stochastic distribution samples δ, Laplace noise $noise$

Output: Global model M

1: Participant downloads round t's global model M^t from the aggregation server;
2: Participants train the VAE model using the raw data $D_{(i)}$ and stochastic distribution samples δ;
3: Participants use the trained VAE to generate forged data $D^f_{(i)}$;
4: Differential privacy is applied to the forged data by participants with Laplace noise $noise$;
5: Each participants trains a local model ${}^f m^t_{(i)}$ using the forged data with differential privacy $D^{f'}_{(i)}$;
6: The aggregation server selects n participants to upload their completed training local models ${}^f m^t_{(i)}$;
7: The aggregation server aggregates the models uploaded by the participants to generate the global model M^{t+1} in round $t+1$;
8: **if** End conditions of federated learning are met **then**
9: Output the global model M;
10: **end if**

V. Experiment

A. Experiment Setup

This section comprehensively evaluates the proposed scheme through the scientific computing libraries Tensorflow and scikit-Learn [19] in python. The experimental environment is configured on the computer of Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz and RTX2060 6G, and the version of Tensorflow used is 2.2.0; the version of scikit-Learn is 0.23.2.

The data used in the simulation are Mnist [20], Fashion_Mnist [21] and Cifar10 [22]. The Mnist is a widely used handwritten digit recognition data, usually used for the performance evaluation of image classification algorithms in the computer vision field. There are 10 number categories in this data, from number 0 to number 9; Fashion_Mnist is an extended version of Mnist, including different types such as T-shirt, Dress, Ankle boot, etc. Cifar10 contains 60,000 color images. All images belong to 10 different categories, such as airplane, dog, truck et al.

B. Experiment Result

We simulate the results of the participants training the VAE generation model. As shown in Fig. 2 (a)(b)(c), as the number of training rounds increases, in different data, the training loss of the generated model VAE gradually decreases until the training model converges. The loss when the Mnist-VAE model converges is about 125, the loss when the Fashion_Mnist-VAE model converges is about 240 and the loss when the Cifar10-VAE model converges is about 1700.

In order to evaluate the feature difference between the raw data generated by VAE and the raw data, we use the same neural network model to train on the forged data and the raw data, generate a forged model and a raw model, and Perform an accuracy test on the test data. As shown in Table. 2, simulations were performed on three different datasets. For different datasets, the models generated by forged data have decreased in predictive accuracy, but they are all within the acceptable range.

TABLE II: Accuracy of Model Trained by Raw and Forged Data

Dataset	Accuracy of Model	
	Trained by Raw Dataset	Trained by Forged Dataset
Mnist	97.6%	95.3%
Fashion_Mnist	91.1%	88.6%
Cifar10	70.6%	67.2%

When performing secure federated learning, participants need to perform local model training and then upload the local model to the aggregation server. In our solution, the participants use the trained vae model to generate forged data, and use this for local training. We evaluated the local training results of forged data on the three datasets. As shown in Fig. 3 and Fig. 4, with the number of training rounds increasing, the local loss and accuracy will continue to converge until the participant ends the local training.

We use forged data to further evaluate the prediction accuracy of the global model under local training. As shown in Fig. 5, we have considered the cases where the number

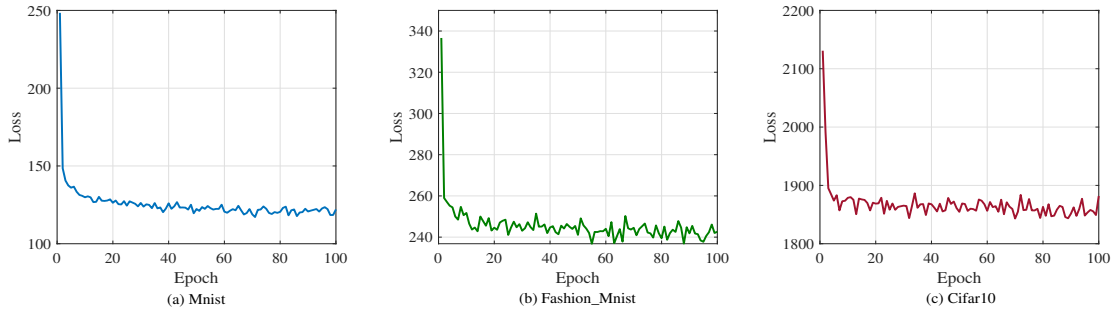

Fig. 2: The Loss of Training VAE.

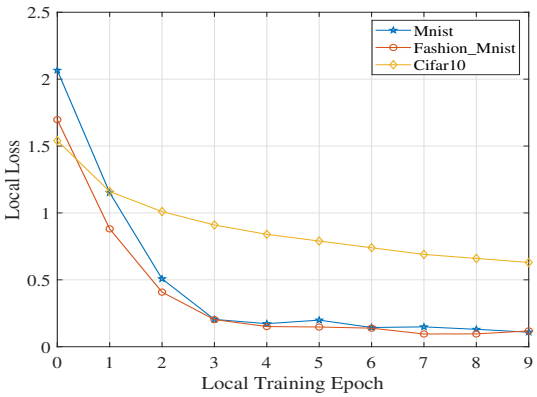

Fig. 3: Participant Local Training Loss.

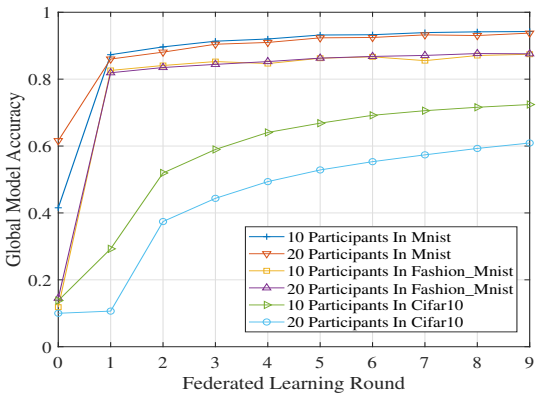

Fig. 5: Global Model Accuracy.

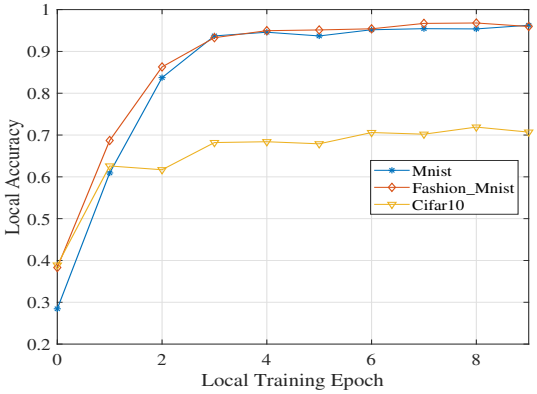

Fig. 4: Participant Local Model Accuracy.

of participants in federated learning is 10 and 20. In a 10-person federated learning system, due to the requirements of the FedAvg algorithm, 5 participants are randomly selected for model aggregation each time. Under the simulations of Mnist, Fashion_Mnist and Cifar10, the prediction accuracy of the global model reached 92%, 87% and 76% respectively. In a 20-person federated learning system, 10 participants are randomly selected for model aggregation each time. Under the simulations of Mnist, Fashion_Mnist and Cifar10, the prediction accuracy of the global model reached 91%, 86% and 64% respectively. Even if participants use forged data for local training, as the number of federated training rounds increases at a high level in different datasets, the accuracy of the global model can still converge to a higher level.

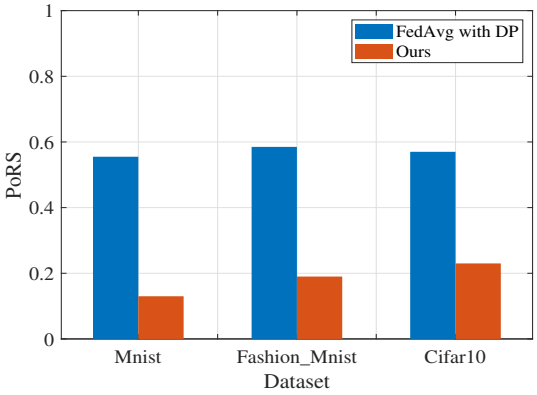

Fig. 6: Probability of Reconstruction Sucess in Different Dataset.

We evaluate the privacy protection of our proposed scheme by reconstructing the probability of success (PoRS). In the literature [23], the author shows that the probability that the opponent reconstructs the training data used by the participant through reasoning is as high as 76%. We compare our approach with the benchmark scheme FedAvg that incorporates differential privacy (DP). Specifically, we believe that the opponent rebuilds the participant's training data at 76%. We perform simulations in three datasets. As shown in Fig. 6, our proposed approach can effectively reduce the probability of reconstruction of the raw data.

VI. Conclusion

In order to reduce the success probability of inference reconstruction attack in federated learning, we proposed a VAE based secure federated learning mechanism. Specifically, participants use raw data to generate forged data through a VAE and train a local model with forged data, thereby protecting the data privacy and guaranteeing the quality of the global model.Experimental results show that the proposed secure federated learning mechanism not only ensures the accuracy of the global model, but also reduces the probability of the successful inference and reconstruction of participants' raw data.

References

[1] J. Mills, J. Hu and G. Min, "Multi-Task Federated Learning for Personalised Deep Neural Networks in Edge Computing," IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 3, pp. 630-641, 2022.

[2] L. Zhao, G. Han, Z. Li and L. Shu, "Intelligent Digital Twin-Based Software-Defined Vehicular Networks," IEEE Network, vol. 34, no. 5, pp. 178-184, 2020.

[3] M. Al-Rubaie and J. M. Chang, "Privacy-Preserving Machine Learning: Threats and Solutions," *IEEE Security & Privacy*, vol. 17, no. 2, pp. 49-58, 2019.

[4] J. Konečnỳ, H. B. McMahan, F. X.Yu, P. Richtárik, A. T. Suresh, D. Bacon,"Federated learning: Strategies for improving communication efficiency," *arXiv preprint* arXiv:1610.05492, 2016.

[5] L. Zhao, W. Zhao, A. Hawbani, A. Al-Dubai, G. Min, A. Y. Zomaya, C. Gong, "Novel Online Sequential Learning-based Adaptive Routing for Edge Software-Defined Vehicular Networks," IEEE Transactions on Wireless Communications, 2020, DOI:10.1109/TWC.2020.3046275.

[6] J. Mills, J. Hu and G. Min, "Communication-Efficient Federated Learning for Wireless Edge Intelligence in IoT," IEEE Internet of Things Journal, vol. 7, no. 7, pp. 5986-5994, 2020.

[7] L. Zhu and S. Han, "Deep leakage from gradients," *Federated Learning*. Springer, Cham, pp. 17-31, 2020.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes",*arXiv preprint arXiv:1312.6114*, 2013.

[9] S. Truex, L. Liu, K. H. Chow, M. E. Gursoy, and W. Wei, "LDP-Fed: Federated learning with local differential privacy," *In Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, 2020, pp. 61-66.

[10] R. Hu, Y. Guo, H. Li, Q. Pei and Y. Gong, "Personalized Federated Learning With Differential Privacy," IEEE Internet of Things Journal, vol. 7, no. 10, pp. 9530-9539, 2020.

[11] N. Mohammadi, J. Bai, Q. Fan, Y. Song, Y. Yi and L. Liu, "Differential Privacy Meets Federated Learning under Communication Constraints," *arXiv preprint arXiv:2101.12240*, 2021.

[12] O. Ibitoye, M. O. Shafiq and A. Matrawy, "DiPSeN: Differentially Private Self-normalizing Neural Networks For Adversarial Robustness in Federated Learning," *arXiv preprint arXiv:2101.03218*, 2021.

[13] S. Kumar, S. Dutta, S. Chatturvedi and M. Bhatia, "Strategies for Enhancing Training and Privacy in Blockchain Enabled Federated Learning," *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 333-340.

[14] X. Luo and X. Zhu, "Exploiting defenses against GAN-based feature inference attacks in federated learning," *arXiv preprint arXiv:2004.12571*, 2020.

[15] B. Liu, Y. Guo and X. Chen, "PFA: Privacy-preserving Federated Adaptation for Effective Model Personalization," *arXiv preprint arXiv:2103.01548*, 2021.

[16] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, J. Joshi and H. Ludwig, "FedV: Privacy-Preserving Federated Learning over Vertically Partitioned Data," *arXiv preprint arXiv:2103.03918*, 2021.

[17] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," *in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, Lauderdale, FL, USA*, 2017, pp. 1273-1282.

[18] H. Hu, Z. Salcic, G. Dobbie and X. Zhang, "Membership Inference Attacks on Machine Learning: A Survey," *arXiv preprint arXiv:2103.07853*, 2021.

[19] A. Jain, "Scikit-learn: Machine learning in Python," *Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.

[20] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *in Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[21] H. Xiao, K. Rasul and R. Vollgraf, "Fashion-Mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[22] Krizhevsky A, Hinton G. "Learning multiple layers of features from tiny images," 2009.

[23] J. Chen, J. Zhang, Y. Zhao, H. Han, K. Zhu and B. Chen, "Beyond Model-Level Membership Privacy Leakage: an Adversarial Approach in Federated Learning," *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, 2020, pp. 1-9.