M-Door: Joint Attack of Backdoor Injection and Membership Inference in Federated Learning

Ye Liu[†], Shan Chang[*], Denghui Li[*], Minghui Dai[*]
[*]School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China
[†]School of Computer Science and Technology, Donghua University, Shanghai, China
006012@jxnu.edu.cn, changshan@dhu.edu.cn, 2222731@mail.dhu.edu.cn, minghuidai@dhu.edu.cn

GLOBECOM 2024 - 2024 IEEE Global Communications Conference | 979-8-3503-5125-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/GLOBECOM52923.2024.10901593

***Abstract*—Federated learning (FL) collaboratively trains global models while preserving private data locally, making it an ideal privacy-preserving learning technique. However, recent studies have shown that FL poses risks of security attacks and privacy leaks during model parameter transfer. Existing research suggests that backdoor attacks cannot assist with membership inference attacks in machine learning. This paper proposes a joint attack of backdoor injection and membership inference in FL, M-Door, which can connect two independent work lines to ensure the security and privacy of FL. In M-Door, an attacker hidden within the client can not only perform backdoor attacks on the global model, but also perform membership inference attacks on the global model by analyzing the transmitted model parameters. This attack method can improve the success rate of backdoor attacks, and simultaneously increase the success rate of membership inference attacks. We conduct extensive experiments on three image classification tasks to evaluate the effectiveness of M-Door. Compared with the other two attack methods, the experimental results show that M-Door exhibits significant advantages in backdoor and membership inference attacks under both IID and Non-IID data settings.**

***Index Terms*—Federated learning, backdoor attacks, membership inference attacks.**

I. Introduction

Federated learning (FL) reduces the risk of sensitive data leakage by retaining data on local devices and only sharing updates to model parameters, while also adhering to the principle of privacy protection [1]. Although FL has been recognized as an effective solution to address data silos and privacy protection, its own security and privacy protection issues have always been a concern for both industry and academia [2] [3].

Currently, two independent long-term work lines study security and privacy issues in FL. From a security perspective, the attacker aims to cause the target machine learning model to behave improperly. Among them, attackers can use backdoor attacks to reduce the performance of the model on target tasks while maintaining good performance on main tasks [4]. From a privacy perspective, the attacker aims to infer private information about the target model or its training data. Among them, the attacker uses a membership inference model to determine whether the given data belongs to the training set of the target model based on its predictions [5].

In recent years, there have been some privacy attacks based on poisoning attacks in machine learning. Mahloujifar *et al.* [6] discuss how poisoning is beneficial for attribute inference attacks, which can infer the attributes of the training dataset. Tramer *et al.* [7] propose a poisoning assisted membership inference attack, which amplifies privacy leakage by injecting poisoned samples into the dataset. Chen *et al.* [8] evaluate the membership inference attacks caused via clean label poisoning. In FL, Wang *et al.* [9] discuss improving attribute inference attacks through data poisoning. However, these papers do not consider the backdoor attacks to enhance privacy inference attacks. Regarding backdoor attacks, Hu *et al.* [10] combine the backdoors with membership inference in machine learning, with the aim of using membership inference to infer whether the model owner is embedded in the backdoor. Liu *et al.* [11] strengthen the single backdoor of early injection by utilizing model information leakage in FL. Goto *et al.* [12] point out that the backdoor assists membership inference attacks because it cannot distinguish the loss distribution between member and non-member samples. To the best of the authors' knowledge, the backdoor attacks in enhancing membership inference attacks in FL have not been studied yet.

Although the above related works simultaneously focus on privacy and security, there are still some weaknesses in their works. Firstly, these studies all borrow one attack to enhance or defend against another attack, and the purpose cannot simultaneously improve the performance of both attacks. Secondly, most of the research scenarios are in machine learning. Since FL is a dynamic scenario, clients can join and exit during FL, making it more challenging in FL scenarios. Finally, security attacks all involve data poisoning through adversarial sample or label flipping, and an increase in the proportion of poisoning will increase the negative impact on normal models, making it easy to detect. In contrast, since the backdoors can be indefinitely hidden until activated by samples with specific backdoor triggers, it poses serious security risks.

To address the above mentioned issues, this work proposes M-Door, a joint attack of backdoor injection and membership inference in FL. The goal of M-Door is to continuously learn triggers that rely on the global model and membership inference attack model based on each round of updated global model and membership inference attack model, while improving the performance of backdoor attacks and membership inference attacks without affecting the performance of the global model. We conduct a large number of experiments to test the performance of M-Door on three image classification

Shan Chang and Minghui Dai are the corresponding authors.

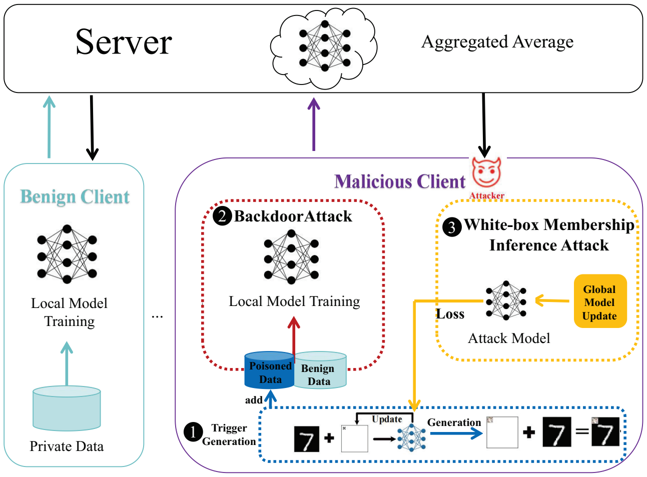


Fig. 1. The Framework of M-Door.

tasks. Compared with the other two attack methods, the results show that M-Door performs superior backdoor attacks and membership inference attacks in all settings.

II. Threat Model

Attack Model: In the FL system, we consider that there is one malicious client (attacker) among N clients. Among them, both the server and the benign client faithfully adhere to the FL protocol, while the malicious client simultaneously implements backdoor attacks and membership inference attacks.

Attacker's Capability: The attacker's capability can be summarized as follows. Firstly, the attacker fully understands the structure and parameters of the global model transmitted by the server. Secondly, the attacker can have complete control over local training on their device. Finally, since we are referring to the white box membership inference attack proposed in the literature [13], we consider that the attacker has an auxiliary dataset and knows some data from other client training datasets. In practical scenarios, attackers can obtain data by participating in other collaborations. In simulation scenarios, attackers know the data format to train the global model and can collect the required data from public datasets.

Attacker's Goals: The attacker aims to achieve the following objectives. (1) Main task effectiveness: ensure the accuracy of the global model on clean samples to prevent it from being discarded. (2) Effectiveness of backdoor attacks: embedding a backdoor into the global model to output specific labels on data with triggers, while not affecting the inference of normal data. (3) Effectiveness of membership inference attacks: conduct a membership inference attack on the global model, i.e., infer whether a certain data sample is in the training dataset of other clients.

III. Design of M-Door

A. Design Overview

In the FL system, an attacker who is hidden within the client and possesses both local and auxiliary datasets can perform backdoor attacks on the global model by attacking the auxiliary dataset, and perform membership inference attacks on the global model by analyzing the model parameters passed through the auxiliary dataset. Please note that in this paper, attackers of membership inference attacks infer whether data records are used to train global models, but do not infer whether data records are used to train specific local models.

Fig. 1 shows the attack process of M-Door. In the current round, after obtaining the global model issued by the server, the malicious client trains the local model normally on the local dataset using the gradient descent method. If an attack is carried out, it performs the following steps.

- **Step 1: Trigger Generation.** Fixed model parameters, based on the current model and auxiliary dataset, using the random gradient ascent algorithm and the loss value of the previous round of membership inference attack model, training and optimizing trigger pixels as parameters to generate the most suitable trigger for the attack requirements in the current round.
- **Step 2: Backdoor Model Training.** Inject triggers into a subset of auxiliary datasets (i.e., the size determined by the poisoning rate) and modify their labels to make them toxic auxiliary data. Combine the poisoning auxiliary data and other benign auxiliary data, train the current model again using the gradient descent method, and send the poisoning model to the server.
- **Step 3: White-box Membership Inference Model Training.** Observe the global model from the previous round, train the membership inference attack model locally based on the auxiliary dataset, and record the loss value after the current round of attack model training, so that it can be used to generate triggers in the next round.

B. Design Principles

Table I shows the factors that affect the two attack methods. It can be observed that there is no correlation between the influencing factors of these two attack methods, which makes it very challenging to find a simple method while improving the success rate of both attacks.

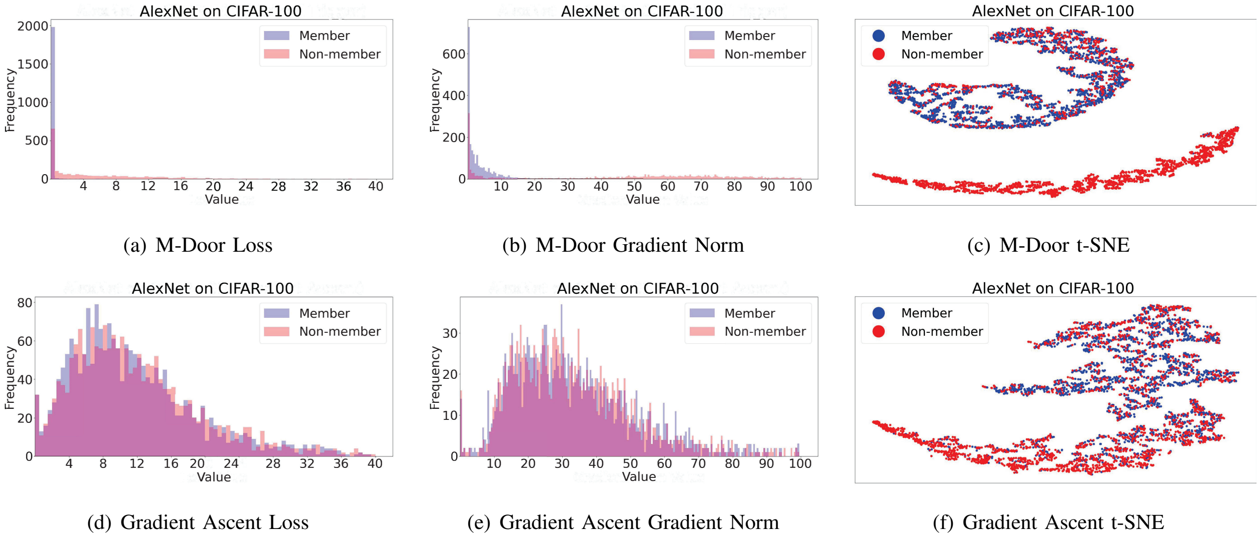
Membership inference attacks are statistical tests with loss distributions [14]. Literature [12] points out that the difference in loss distribution between member and non-member data is also essential for amplifying privacy leaks. As backdoors cannot distinguish the loss distribution between member and non-member samples, backdoor attacks cannot assist membership inference attacks.

For backdoor attacks, the success rate largely depends on designing specific triggers based on the characteristics of the target task. So, we speculate that if a trigger can be trained to distinguish the loss distribution between member and non-member samples, the backdoor can assist in membership inference attacks.

Therefore, in M-Door, we link backdoor and membership inference attacks by designing a loss function generated by triggers. Specifically, the loss function for generating triggers is divided into two parts. One part is to reduce the probability

TABLE I
THE INFLUENCING FACTORS OF BOTH ATTACKS

Attack Methods	**Influence Factor**
Backdoor Attack	Trigger properties; Poisoning rate.
Membership Inference Attack	Overfitting of the target model; The type of target model; Diversity of target model training data.



(a) M-Door Loss

(b) M-Door Gradient Norm

(c) M-Door t-SNE

(d) Gradient Ascent Loss

(e) Gradient Ascent Gradient Norm

(f) Gradient Ascent t-SNE

Fig. 2. The Difference Between M-Door and Gradient Ascent.

of the corresponding sample label, thereby ensuring the effectiveness of backdoor attacks. The other part is the loss value of the membership inference attack model, which enables the generated trigger to learn knowledge about the loss distribution of distinguishing member samples from non-member samples in the membership inference attack model, thereby assisting membership inference attacks.

C. Trigger Generation and Model Training

(1) Trigger Generation. In M-Door, the generated trigger needs to meet the following two objectives: improve the effectiveness of backdoor attacks and distinguish the loss distribution between member and non-member samples. Therefore, M-Door refines the loss function generated by the trigger to generate a trigger t that meets the attack requirements.

Backdoor loss: To ensure the effectiveness of backdoor attacks, the trained model must output the correct class label probability as small as possible when encountering samples with triggers.

$$L_{backdoor} = \max_{t} L(w,t) = \frac{1}{K}\sum_{i=1}^{K} L(w, x_i + t), \qquad (1)$$

where w represents the neural network weights, $x_i, i = 1, ..., K$ represents the training samples, and $L(\cdot)$ is the loss function. The $\max(.)$ function in Eq. (1) represents finding a set of backdoor samples in the sample space that maximizes the loss function.

Membership inference loss: In order to enable the generated trigger to distinguish the loss distribution between member and non-member samples, the loss value of the membership inference attack model A trained in the previous round is expressed as

$$L_{membership} = L_{attack}\left(A\left(x; w\right), y\right). \qquad (2)$$

In order to limit the disturbance range of the trigger and ensure the low visibility and smoothness generated by the trigger, t is projected onto the l_p-norm sphere to prevent it from being too large. ϵ controls the l_p diameter of the bounded perturbation.

$$\|t\|_p \leq \epsilon. \qquad (3)$$

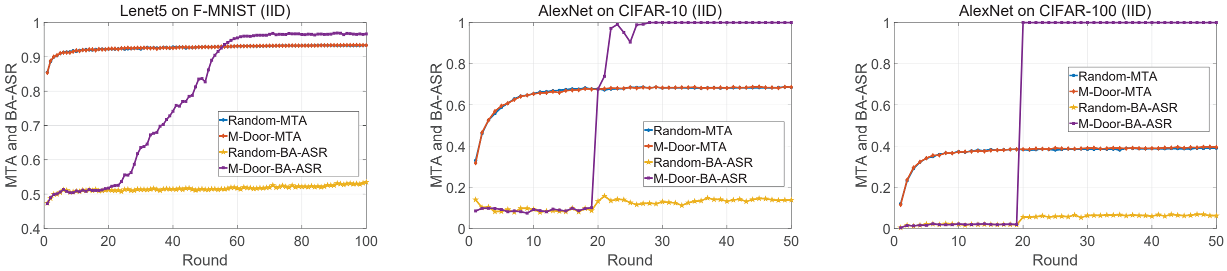
In summary, the loss function generated by the designed trigger is

$$\begin{cases} L_t = L_{backdoor} + L_{membership}, \\ \qquad s.t.\ \|t\|_p \leq \epsilon. \end{cases} \qquad (4)$$

According to the loss function, M-Door learns the most suitable backdoor trigger t based on the current global model and membership inference attack model.

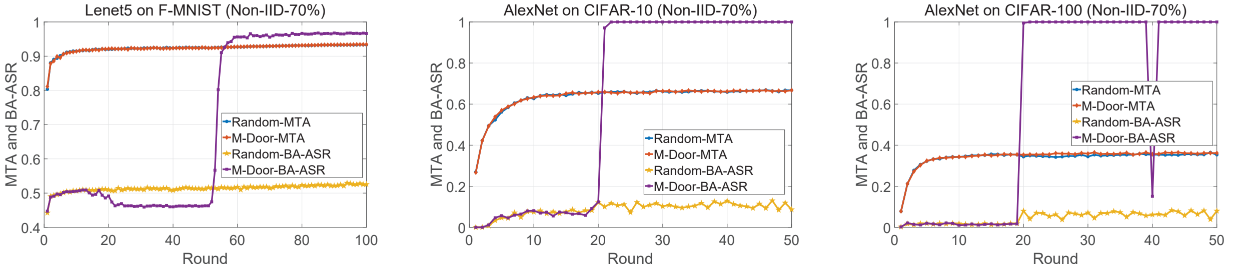
(2) Backdoor Model Training. After obtaining the generated trigger t, inject the trigger into a portion of the auxiliary dataset and modify its label to make it toxic auxiliary data. Then, combine the poisoning auxiliary data with benign auxiliary data, train the current model based on the gradient descent method, and upload it to the server.

(3) White-box Membership Inference Model Training. In recent years, one of the most extensively studied works is the white-box membership inference attack proposed by Nasr *et al.* [13]. They conduct active and passive membership inference attacks in different threat scenarios. We refer to



(a) CNN on F-MNIST (IID)   (b) AlexNet on CIFAR-10 (IID)   (c) AlexNet on CIFAR-100 (IID)

Fig. 3. Backdoor Attack Performance of Various Methods under Different Tasks (IID).



(a) CNN on F-MNIST (Non-IID-70%)   (b) AlexNet on CIFAR-10 (Non-IID-70%)   (c) AlexNet on CIFAR-100 (Non-IID-70%)

Fig. 4. Backdoor Attack Performance of Various Methods under Different Tasks (Non-IID-70%).

TABLE II
NORMA MODEL PARAMETER

Tasks	CNN on F-MNIST	AlexNet on CIFAR-10	AlexNet on CIFAR-100
Optimizer	SGD	SGD	SGD
Batch Size	64	64	64
Local Epoch	3	5	10
Total Number of Clients	15	10	5
Participation Rate	100%	100%	100%

passive membership inference attack methods and compare it with the active membership inference attack.

IV. EXPERIMENTAL

A. Experimental Setup

Models and Dataset: For the dataset, we use three universal datasets: F-MNIST, CIFAR-10, and CIFAR-100. For the model structure, we use the F-MNIST datasets to train a six-layer CNN model, and use the CIFAR-10 and CIFAR-100 datasets to train an eight-layer AlexNet model.

Parameter Settings: In the experiment, when the global model was trained normally, the learning rates for all tasks were 0.01, 0.001, and 0.0001 for rounds 0-50, 50-100, and 100-200, respectively. The remaining model parameter settings are shown in Table II. When the attacker carries out active attacks (backdoor attacks and Gradient Ascent), the attack parameter settings are shown in Table III. Please note that if the Gradient Ascent attacks too many times, it will cause the global model to fail. Therefore, the Gradient Ascent in this experiment implements MIA-ASR while ensuring MTA.

When the attacker implements a membership inference attack, in all tasks, the batch size is 64; the learning rate is 0.0002, and the optimizer is Adam. The training and testing data of the membership inference attack model includes member and non-member data, where member data is randomly extracted from the entire training dataset without being placed back, and non-member data is randomly extracted from the entire testing set without being placed back. The dataset used for membership inference attacks and active attacks is the same batch of data. Since the maximum effectiveness of backdoor and membership inference attacks is usually achieved when the FL model converges, we choose to attack after the model converges.

Comparison Algorithm: We compare the following three algorithms:

- *Gradient Ascent:* Attackers implement gradient ascent for active membership inference attacks [13].
- *Random:* Attackers use random triggers for backdoor attacks and membership inference attacks.
- *M-Door:* Attackers use trained triggers for backdoor attacks and membership inference attacks.

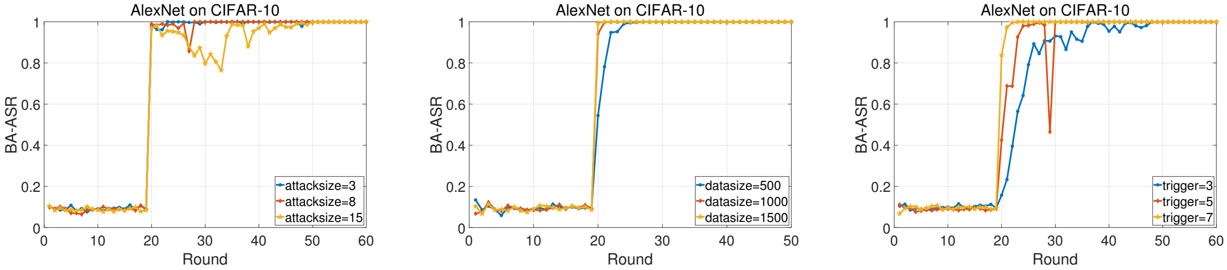
Evaluation Metrics: We use three metrics, Main Task Accuracy (MTA), Backdoor Attack Success Rate (BA-ASR), and Membership Inference Attack Success Rate (MIA-ASR) to evaluate the overall performance of M-Door.

- *MTA:* The prediction accuracy of the global model on the test dataset.
- *BA-ASR:* Measures the effectiveness of backdoor attacks. Calculate the percentage of malicious input errors classified to the target label through a backdoor model.
- *MIA-ASR:* The success rate of membership inference attacks is the score of the correct member prediction of unknown data.

TABLE III
ATTACK MODEL PARAMETER

Tasks	CNN on F-MNIST	AlexNet on CIFAR-10	AlexNet on CIFAR-100
Optimizer	Adam	Adam	Adam
Batch Size	3	64	64
Local Epoch	3	5(Backdoor),1(Gradient Ascent)	10(Backdoor),1(Gradient Ascent)
Learning Rate	0.00002	0.00001	0.000001
Trigger Size	2×2	10×10	10×10
Start Attack Round	10	20	20
***Auxiliary Dataset Size**	(1500,1500,1500,1500)	(2000,2000,2000,2000)	(2500,2500,2500,2500)
Poisoning Rate (Train, Test)	(0.01,0.5)	(1,1)	(1,0.5)

*Auxiliary Dataset Size (Train Member, Train Non-Member, Test Member, Test Non-Member)



(a) Impact of attack interval frequency

(b) Impact of auxiliary dataset size

(c) Impact of trigger size

Fig. 5. Impact of Different Parameter Settings on Backdoor Attacks.

B. Validity Explanation

In this subsection, we discuss from three perspectives why the MIA-ASR of M-Door is higher than that of Gradient Ascent. Therefore, we separately analyze the differences in loss distribution [14], gradient norm [13], and t-SNE [15] of the target model between M-Door and Gradient Ascent.

The Difference in Loss Distribution: Fig. 2(a) and Fig. 2(d) show the loss distribution of member and non-member samples in the M-Door attack and Gradient Ascent attack in the AlexNet on CIFAR-100 task. It can be observed that for the loss distribution of member and non-member samples, the M-Door attack is clearly separated, while the Gradient Ascent attack is not clearly separated.

The Difference in Gradient Norm: Fig. 2(b) and Fig. 2(e) show the distribution of the last layer gradient specification for member and non-member samples in the M-Door attack and Gradient Ascent attack in the AlexNet on CIFAR-100 task. It can be observed that for the gradient norm distribution of member and non-member samples, M-Door attacks are easy to distinguish, while Gradient Ascent attacks are difficult to distinguish.

The Impact of M-Door and Gradient Ascent on t-SNE: Fig. 2(c) and Fig. 2(f) show the output of member and non-member samples after clustering in the last second layer of the attack model in the M-Door attack and Gradient Ascent attack in the AlexNet on CIFAR-100 task. It can be observed that for the t-SNE distribution of member and non-member samples, M-Door attacks are easy to distinguish, while Gradient Ascent attacks are difficult to distinguish.

C. Experimental Results

This experiment shows the MTA, BA-ASR, and MIA-ASR of different methods under IID and Non-IID-70% data distribution in three different tasks (CNN on F-MNIST, AlexNet on CIFAR-10, and AlexNet on CIFAR-100).

Backdoor Attack: Fig. 3 (IID) and Fig. 4 (Non-IID-70%) show MTA and BA-ASR convergence trends for each attack method under IID and Non-IID-70% data distributions, respectively. Experimental results show that regardless of the situation, the BA-ASR of M-Door is higher than that of Random, and the MTA is not affected, indicating the effectiveness of M-Door backdoor attacks.

Membership Inference Attack: Table IV shows the MTA, BA-ASR, and MIA-ASR of each attack method for each task under IID and Non-IID-70% data distribution. Experimental results show that compared to other attacks, M-Door can achieve high-precision MIA-ASR, validating the effectiveness of M-Door membership inference attacks.

D. Ablation Study

In this subsection, the effects of different parameter settings, such as attack interval frequency, auxiliary dataset size, and trigger size on MTA, BA-ASR, and MIA-ASR are analyzed.

Impact of Attack Interval Frequency: In the AlexNet on CIFAR-10 task, with other parameters unchanged, observe the BA-ASR and MIA-ASR of the M-Door method for different attack intervals (3, 8, 15). Fig. 5(a) shows that the number of attack intervals slightly impacts the backdoor attack process, but ultimately reaches 100% BA-ASR. Table V shows that as the number of attack intervals increases, MTA improves but MIA-ASR decreases.

TABLE IV
MEMBERSHIP INFERENCE ATTACK PERFORMANCE OF VARIOUS METHODS UNDER DIFFERENT TASKS

Tasks	Metrics	IID			Non-IID-70%		
		Gradient Ascent	Random	M-Door	Gradient Ascent	Random	M-Door
CNN on F-MNIST	MTA	90.08%	93.29%	93.29%	90.39%	93%.13	93.17%
	BA-ASR	-	52.13%	96.51%	-	52.33%	96.65%
	MIA-ASR	52.87%	53.90%	54.27%	66.56%	67.42%	67.66%
AlexNet on CIFAR-10	MTA	64.54%	67.23%	67.96%	63.66%	65.73%	66.28%
	BA-ASR	-	21.35%	100 %	-	16.61%	100%
	MIA-ASR	66.45 %	70.45%	71.30%	78.25%	81.00%	81.35%
AlexNet on CIFAR-100	MTA	39.07%	37.94%	38.98%	36.29%	35.33%	36.16%
	BA-ASR	-	8.69%	100%	-	7.98%	100%
	MIA-ASR	78.86%	79.88%	79.90%	78.86%	78.61%	78.90%

TABLE V
IMPACT OF DIFFERENT PARAMETER SETTINGS ON MEMBERSHIP INFERENCE ATTACKS

Parameter	Values	MTA	BA-ASR	MIA-ASR
Attack Interval Frequency	3	67.83%	100%	71.83%
	8	67.64%	100%	70.55%
	15	67.93%	100%	70.53%
Auxiliary Dataset Size	500	68.05%	100%	70.90%
	1000	67.74%	100%	71.20%
	1500	67.76%	100%	71.10%
Trigger Size	3×3	67.67%	100%	70.40%
	5×5	67.84%	100%	70.43%
	7×7	67.67%	100%	71.30%

Impact of Auxiliary Dataset Size: In the AlexNet on CIFAR-10 task, with other parameters remaining unchanged, observe the BA-ASR and MIA-ASR of the Door Member method for different auxiliary dataset sizes (500,1000,1500). Fig. 5(b) shows that even with a very small auxiliary dataset, the final BA-ASR will reach 100%. Table V shows that the larger the auxiliary dataset, the lower the MTA and the higher the MIA-ASR.

Impact of Trigger Size: In the AlexNet on CIFAR-10 task, with other parameters unchanged, observe the BA-ASR and MIA-ASR of the M-Door method for different trigger sizes (3×3, 5×5, 7×7). Fig. 5(c) shows that even with a very small trigger, the final BA-ASR will reach 100%. Table V shows that trigger size has almost no effect on MTA, but MIA-ASR improves when the trigger becomes larger.

V. CONCLUSION

In this paper, we have proposed a novel method, M-Door, a joint attack of backdoor injection and membership inference in FL. Specifically, we have designed a loss function for generating triggers, which partly reduces the probability of the corresponding label of the sample to ensure the effectiveness of backdoor attacks, and partly measures the loss value of the membership inference attack model to ensure the effectiveness of the membership inference attack. In M-Door, triggers are generated based on the global model and membership inference attack model, enhancing backdoor and membership inference attacks. Through extensive experiments, we have demonstrated the effectiveness of M-Door.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundation of Shanghai (Grant No. 22ZR1400200), the Fundamental Research Funds for the Central Universities (No. 2232023Y-01), the National Natural Science Foundation of China (Grant No. 62472083).

REFERENCES

[1] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, "A survey on federated learning: challenges and applications," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 513–535, 2023.

[2] T. Nguyen and M. T. Thai, "Preserving privacy and security in federated learning," *IEEE/ACM Transactions on Networking*, 2023.

[3] M. Dai, T. Wang, Y. Li, Y. Wu, L. Qian, and Z. Su, "Digital twin envisioned secure air-ground integrated networks: A blockchain-based approach," *IEEE Internet of Things Magazine*, vol. 5, no. 1, pp. 96–103, 2022.

[4] T. Liu, Y. Zhang, Z. Feng, Z. Yang, C. Xu, D. Man, and W. Yang, "Beyond traditional threats: A persistent backdoor attack on federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 21359–21367, 2024.

[5] G. Zhu, D. Li, H. Gu, Y. Han, Y. Yao, L. Fan, and Q. Yang, "Evaluating membership inference attacks and defenses in federated learning," *arXiv preprint arXiv:2402.06289*, 2024.

[6] S. Mahloujifar, E. Ghosh, and M. Chase, "Property inference from poisoning," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1120–1137, 2022.

[7] F. Tramèr, R. Shokri, A. San Joaquin, H. Le, M. Jagielski, S. Hong, and N. Carlini, "Truth serum: Poisoning machine learning models to reveal their secrets," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2779–2792, 2022.

[8] Y. Chen, C. Shen, Y. Shen, C. Wang, and Y. Zhang, "Amplifying membership exposure via data poisoning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29830–29844, 2022.

[9] Z. Wang, Y. Huang, M. Song, L. Wu, F. Xue, and K. Ren, "Poisoning-assisted property inference attack against federated learning," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[10] H. Hu, Z. Salcic, G. Dobbie, J. Chen, L. Sun, and X. Zhang, "Membership inference via backdooring," *arXiv preprint arXiv:2206.04823*, 2022.

[11] T. Liu, X. Hu, and T. Shu, "Technical report: Assisting backdoor federated learning with whole population knowledge alignment," *arXiv preprint arXiv:2207.12327*, 2022.

[12] Y. Goto, N. Ashizawa, T. Shibahara, and N. Yanai, "Do backdoors assist membership inference attacks?," *arXiv preprint arXiv:2303.12589*, 2023.

[13] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753, 2019.

[14] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022.

[15] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.