FedMIA: An Effective Membership Inference Attack Exploiting "All for One" Principle in Federated Learning

Gongxi Zhu[1] Donghao Li[2] Hanlin Gu[2,3] * Yuan Yao[2] Lixin Fan[3] Yuxing Han[1]
[1]Tsinghua University [2]The Hong Kong University of Science and Technology [3]Webank
gx.zhu@foxmail.com, dlibf@connect.ust.hk, ghltsl123@gmail.com

Abstract

Federated Learning (FL) is a promising approach for training machine learning models on decentralized data while preserving privacy. However, privacy risks, particularly Membership Inference Attacks (MIAs), which aim to determine whether a specific data point belongs to a target client's training set, remain a significant concern. Existing methods for implementing MIAs in FL primarily analyze updates from the target client, focusing on metrics such as loss, gradient norm, and gradient difference. However, these methods fail to leverage updates from non-target clients, potentially underutilizing available information. In this paper, we first formulate a one-tailed likelihood-ratio hypothesis test based on the likelihood of updates from non-target clients. Building upon this formulation, we introduce a three-step Membership Inference Attack (MIA) method, called FedMIA, which follows the "all for one"—leveraging updates from all clients across multiple communication rounds to enhance MIA effectiveness. Both theoretical analysis and extensive experimental results demonstrate that FedMIA outperforms existing MIAs in both classification and generative tasks. Additionally, it can be integrated as an extension to existing methods and is robust against various defense strategies, Non-IID data, and different federated structures. Our code is available in https://github.com/Liar-Mask/FedMIA.

1. Introduction

Federated learning (FL) [18, 25, 26] has emerged as a promising approach for training machine learning models on decentralized data sources while ensuring data privacy. Despite its advantages, the privacy risks associated with the information exchanged during FL have attracted significant research attention. Membership Inference Attacks (MIAs) in FL aim to determine whether a specific data point was part of a particular client's training dataset, typically performed by adversaries positioned on the server side. In contrast to Gradient Inversion Attacks (GIAs) [7, 46], MIAs [27] do not rely on strong assumptions, such as small batch sizes or local training epochs, and thus remain significantly underexplored within the FL context.

*Corresponding author.

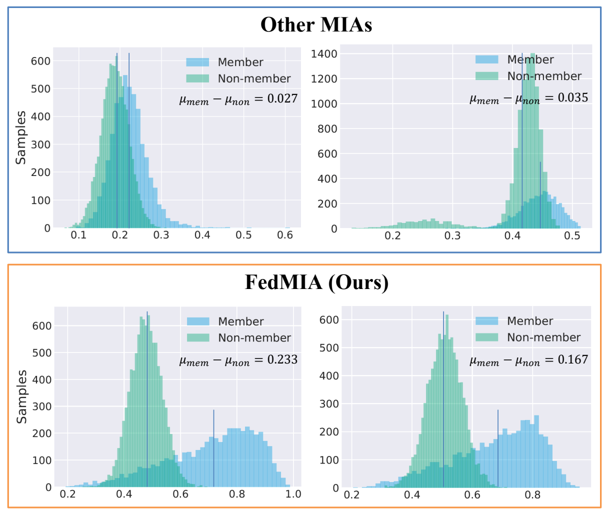


Figure 1: The distributions of member and non-member samples of FedMIA (the second row: FedMIA-I (ours) and FedMIA-II (ours)) and other MIAs (the first row: Grad-Cosine [23], Loss-Series [10]) on ResNet-CIFAR100. It shows the obvious gap between the mean of the member and non-member ($\mu_{mem} - \mu_{non}$) for the proposed FedMIA compared to other methods.

Most existing MIAs in FL [23, 27, 28, 36, 41, 44] focus on inferring membership *solely from the updates of the target clients*, utilizing gradient norms, loss values, and gradient differences. However, these methods overlook the valuable information contained in updates uploaded by non-target clients. Recent work [10, 13] attempts to enhance the effectiveness of MIAs by incorporating shadow models into FL. While these methods make use of additional information, they require an auxiliary dataset to train the shadow model, which may not be feasible in the FL setting, as the server does not have access to the private data of local clients. Furthermore, training a shadow model incurs additional computational costs for the adversary.

To address the limitations of existing approaches, this paper proposes an alternative method that leverages updates from non-target clients, denoted as $I_{\text{non-tar}}$, instead of relying on an auxiliary dataset for Membership Inference Attacks (MIAs). A key challenge in this approach is that the server lacks knowledge of which updates correspond to data trained on the target client's dataset, making it difficult to estimate the distribution of updates trained on the target data, denoted as $\mathcal{Q}_{\text{in}}$. To overcome this challenge, we demonstrate that it is possible to estimate the distribution of updates that are not trained on the target data, denoted as $\mathcal{Q}_{\text{out}}$, using updates uploaded by non-target clients $I_{\text{non-tar}}$. Since each client's data is disjoint, at least some of the updates from non-target clients $I_{\text{non-tar}}$ will not be trained on the target data, i.e., $\exists I_{\text{non-tar}} \sim \mathcal{Q}_{\text{out}}$. Based on this, we formulate a one-tailed likelihood-ratio hypothesis test using the distribution of updates that were not trained on the target data, $\mathcal{Q}_{\text{out}}$, to perform the MIA.

Building on this one-tailed likelihood-ratio hypothesis test, we introduce a three-step Membership Inference Attack (MIA) method called FedMIA, which follows the "all for one"—leveraging updates from all clients across multiple communication rounds. The first step involves computing a low-dimensional representation to simplify the distribution of updates. In the second step, we estimate the distribution of updates not trained on the target data for each communication round. Finally, we apply the one-tailed likelihood-ratio test based on the estimated distribution of updates not trained on the target data. This test is further extended by incorporating updates from all communication rounds. The proposed FedMIA has three advantages: 1) FedMIA achieves superior performance compared to other MIAs by utilizing the updates information from non-target clients in Sect. 4; 2) FedMIA can be integrated into existing methods as an extension in Sect. 4.1; and 3) FedMIA is robust against six defense methods, two federated structures, varying degrees of Non-IID data, and different client counts, communication rounds, and local epochs. Our contributions are summarized as the following:

- We first formulate the non-target updates as a one-tailed likelihood-ratio hypothesis test to evaluate the performance of the updates without being trained on target data. Theorem 1 proves the validity of our formulation.
- Building on this hypothesis test, we introduce a three-step Membership Inference Attack (MIA) method, called FedMIA, which leverages updates from all clients and communication rounds to enhance MIA effectiveness.
- Extensive results show that the proposed FedMIA: 1) achieves superior performance compared to other MIAs in both classification and generative tasks (see Fig. 1); 2) can be integrated into existing methods as a plug-in; and 3) is robust against six defense methods, two federated structures, varying degrees of Non-IID data, and different client counts, communication rounds, and local epochs.

2. Related work

2.1. Federated Learning

Federated learning was originally proposed as a collaborative approach for training machine learning models without the need to share private data among multiple parties [18, 25, 26, 39]. However, more recently, the concept of "trustworthy federated learning" has been introduced by [17]. This variant of federated learning places a heightened emphasis on the preservation of privacy throughout the federated learning process. This shift in focus reflects the increasing awareness of privacy concerns and the recognition of the importance of robust security measures in federated learning systems

2.2. Membership Inference Attack

MIA is a widely studied privacy attack in centralized learning scenarios. Depending on the information available to the attacker, MIA can be categorized into black-box attack (where only the output of the model can be obtained) [4, 16, 31–33, 35, 37, 40] and white-box attack (where the entire model is available) [28, 29].

In the context of federated learning, Nasr et al. [28] first analyzed membership inference attacks in federated learning and proposed both passive and active attacks. In a passive attack, the attacker solely focuses on obtaining membership leaks based on accessible information without disrupting or compromising the normal training process. Conversely, an active attack involves the ability to modify the updates of federated learning, thereby increasing the vulnerability of the trained models to attacks. Zari et al. [41] proposed a membership inference attack for federated learning that utilizes the probabilities of correct labels under local models at different epochs for inference. However, this approach requires member samples for auxiliary attacks. Li et al. [23] proposed a passive membership inference attack that does not require training on member samples. They designed two metric features based on the orthogonality of gradients to distinguish whether a sample is a member. Hu et al.[15] designed an inference attack to facilitate an honest-but-curious server to identify the training record's source client, which bases on but extends MIAs to source inference. Moreover, inspired by work on worst-case privacy auditing, Aerni et al. [2] introduced an efficient assessment method that accurately reflects the privacy of defenders at their most vulnerable data points.

3. An Effective MIA in FL

In this section, we first present the setting of federated learning (FL) in Sect. 3.1. Subsequently, we formulate the Membership Inference Attacks (MIAs) in FL as a one-tailed likelihood ratio test in Sect. 3.2. Building upon this formulation, we introduce an effective MIA in Sect. 3.3.

3.1. Setting

Horizontal Federated Learning. We consider a *horizontal federated learning* (HFL) [25, 39] setting consisting of one server and K clients. We assume K clients have their local dataset $D_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k}, k = 1 \cdots K$, where $x_{k,i}$ is the input data, $y_{k,i}$ is the label, and n_k is the total number of data points for k_{th} client. Since we focus on evaluating membership on each client, *we further assume D_k are disjoint*. We consider two commonly-used FL frameworks: 1) FedAvg [26] that the server aggregates the models uploaded by all clients; 2) FedEmbedding [24] that the server aggregates the embeddings uploaded by all clients;

Threat Model. We assume the server are *semi-honest* and do not collude with each other. The server faithfully executes the training protocol but aims to infer the membership information from the specific local clients.

Specifically, the server implement the attack $\mathcal{A}$ determine whether a specific sample (x, y) belongs to the target client's dataset D_{tar} based on a series of updates (models, gradients or embeddings) among K clients and T communication rounds: $\mathcal{I} = \{I_k^t | t \in [T], k \in [K]\}$. However, the server will not actively manipulate these information from the local clients and thus without affecting the utilities of the FL model. The $\mathcal{A}$ can be represented as the following:

$$\mathcal{A}(x, y, \mathcal{I}) = \begin{cases} 1, & \text{if } (x, y) \in D_{tar} \\ 0. & \text{otherwise} \end{cases} \tag{1}$$

3.2. One-tailed Likelihood-Ratio Test in FL

In FL, a membership inference attack is to determine the sample (x, y) belonging to target dataset D_{tar} as follows:

$$H_0 : (x, y) \notin D_{tar}, \qquad H_1 : (x, y) \in D_{tar}. \tag{2}$$

Moreover, for each communication round, the server observes K updates $\{I_k^t\}_{k=1}^K$. According to the updates of the target client I_{tar}^t, it is thus natural to see a membership inference attack as performing to guess whether updates I_{tar}^t is trained on the data (x, y) or not. The Likelihood-ratio Test [3] can be represented as:

$$\Lambda(I_{tar}; x, y) = \frac{p_{in}(I = I_{tar} | x, y)}{p_{out}(I = I_{tar} | x, y)}, \tag{3}$$

where $\mathbb{Q}_{in}(I|x, y)$ and $\mathbb{Q}_{out}(I|x, y)$ denote the distribution of updates trained on datasets with and without (x, y), respectively, and p_{in} and p_{out} represents the probability density function of $\mathbb{Q}_{in}(I|x, y)$ and $\mathbb{Q}_{out}(I|x, y)$.

However, *estimating $\mathbb{Q}_{in}(x, y)$ in federated learning (FL) presents a challenge, as the attacker (e.g., the server) lacks knowledge of which updates are trained on (x, y)*. Therefore, we build the one-tailed Likelihood-Ratio Test using distribution $\mathbb{Q}_{out}(I|x, y)$ as:

$$\hat{\Lambda}(I_{tar}, x, y) = \sum_{I' < I_{tar}} p_{out}(I = I' | x, y), \tag{4}$$

which is the probability of observing a confidence as high as the target updates under the null-hypothesis that the target point (x, y) is a non-member.

Remark 1. *It is noted that, given the data sample (x, y), and considering that clients' training datasets are disjoint, at least $K - 2$ updates (except the target updates) are guaranteed not to be trained on (x, y). This enables the estimation of $\mathbb{Q}_{out}(x, y)$.*

Remark 2. *Eq. (4) assumes that the member corresponds to a large value of I. If the member corresponds to a smaller value of I, Eq. (4) is the probability of observing a confidence as high as the target updates.*

Furthermore, when the server observes the $\mathcal{I} = \{I_k^t | t \in [T], k \in [K]\}$ during T communication rounds, we can extend Eq. (4) as the following by utilizing the temporal information:

$$\begin{aligned} \tilde{\Lambda}(\{I_{tar}^t\}_{t=1}^T, x, y) &= \frac{1}{T} \sum_{t=1}^{T} \hat{\Lambda}(I_{tar}^t, x, y) \\ &= \frac{1}{T} \sum_{t=1}^{T} \sum_{I' < I_{tar}^t} p_{out}(I = I' | x, y), \end{aligned} \tag{5}$$

where $\mathbb{Q}_{out}^t(I|x, y)$ are the distribution of updates trained on datasets without (x, y) in the t_{th} communication round. We establish the validity of Eq. (5) in Theorem 1, which demonstrates that for all T communication rounds, any member inferred by $\tilde{\Lambda}$ is also inferred at least once by $\hat{\Lambda}^t, t \in [T]$. Furthermore, the worst-case membership leakage occurs when the target sample is inferred as a member in at least one communication round (see proof in Appendix C).

Theorem 1. *Given the threshold δ, let $\mathbb{V}_t$ be the member sets estimated by $\hat{\Lambda}^t$ and δ in communication round t. Let $\tilde{\mathbb{V}}$ be the member sets estimated by $\tilde{\Lambda}$ and δ. Then we have*

$$\tilde{\mathbb{V}} \subset (\mathbb{V}_1 \cup \cdots \cup \mathbb{V}_T). \tag{6}$$

3.3. The Proposed Method: FedMIA

Based on the one-tailed likelihood-ration test illustrated in Sect. 3.2, we propose a three-step Membership inference attacks by leveraging the spatial and temporal information.

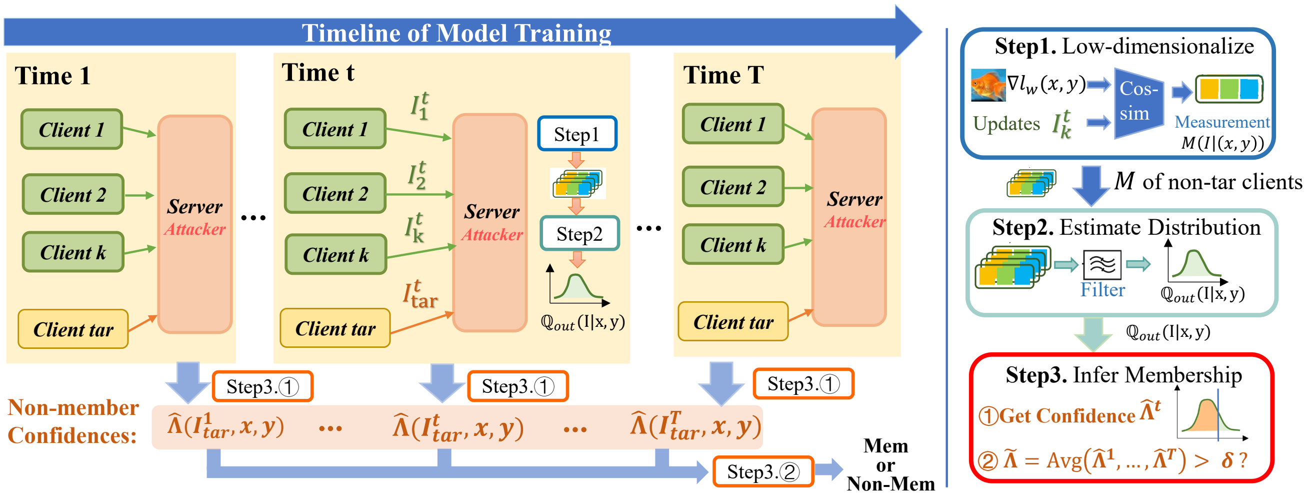


Figure 2: Overview of FedMIA including three steps: 1) Computing the low-dimensional measurement; 2) Estimating the distribution of updates without being trained on target data; 3) Building the one-tailed LRT test and Inferring the membership.

Step 1: Computing the low-dimensional measurement $M(I|(x,y))$.

Since the updates are high-dimensional, directly estimating $\mathbb{Q}_{out}(I|x,y)$ is challenging. To address this, we utilize gradient similarity [23] to map the updates $I(x,y)$ to a low-dimensional variable $M(I|(x,y))$, which allows for an estimation of the distribution as follows:

$$M(I|(x,y)) = \frac{\langle I, \frac{\partial \ell(\omega,x,y)}{\partial \omega}\rangle}{\|I\| \|\frac{\partial \ell(\omega,x,y)}{\partial \omega}\|}, \tag{7}$$

Eq. (7) evaluates the similarity between the uploaded gradient I (∇F) and the model gradient on the target data $\frac{\partial \ell(\omega,x,y)}{\partial \omega}$, with similarity increasing when the target data is a member. The model ω is the global model.

Remark 3. *Eq. (7) represents one approach; other measurements, such as loss or gradient norm [10, 27], can also be used. This suggests that our method can be integrated with existing methods that employ different measurements. In Sect. 4, we also consider the model loss $\ell(\omega,x,y)$ on target data [10] as one measurement.*

Step 2: Estimating the $\mathbb{Q}_{out}(M(I|x,y))$ for each communication round t.

We first leverage the non-target clients' updates $\{I_k^t | k \in [K], k \neq tar\}$ to estimate the distribution of $M(I|(x,y)$ without trained on (x,y), i.e., $\tilde{\mathbb{Q}}_{out}(M(I|(x,y)))$. We assume $\mathbb{Q}_{out}(M(I|x,y))$ is a Gaussian distribution, i.e., $\mathbb{Q}_{out}(M(I|x,y)) \sim \mathcal{N}(\mu_{out}, \sigma_{out}^2)$.

If (x,y) is trained on I_k^t, then $M(I_k^t|x,y)$ becomes large. Therefore, if $M(I_k|(x,y))$ is exceptionally high for all non-target clients' measurements, the updates from the non-target client k are likely trained on (x,y) with high probability. Consequently, we remove the extreme large values of $M(I_k^t|x,y)$, where $k \neq$ tar, to better estimate $\tilde{\mathbb{Q}}_{out}(M(I|(x,y)))$. To filter the updates sets, we apply the 3-σ rule to the updates trained on (x,y). Specifically, we remove the k-th update if the following condition holds:

$$M(I_k^t|x,y) > \mu^t + 3\sigma^t, \tag{8}$$

$$\text{where } \begin{cases} \mu^t = \frac{1}{K-1}\sum_{j\neq \text{tar}} M(I_j^t|x,y) \\ \sigma^t = \sqrt{\frac{1}{K-1}\sum_{j\neq \text{tar}}(M(I_j^t|x,y)-\mu^t)^2}. \end{cases} \tag{9}$$

After filtering, we obtain the update set $\mathcal{U}_t$ which excludes updates trained on the target data (x,y) for communication round t with high probability. Therefore, we estimate the distribution mean and variance of $\mathcal{N}(\mu_{out}^t, v_{out}^t)$ as:

$$\begin{cases} \mu_{out}^t = \frac{1}{|\mathcal{U}^t|}\sum_{j\in\mathcal{U}^t} M(I_j^t|x,y), \text{ and} \\ v_{out}^t = \frac{1}{|\mathcal{U}^t|}\sum_{j\in\mathcal{U}^t}(M(I_j^t|x,y)-\mu_{out}^t)^2. \end{cases} \tag{10}$$

Step 3: Inferring the membership based on $\tilde{\Lambda}(\{I_{tar}^t\}_{t=1}^T, x, y)$.

According the μ_{out}^t, and v_{out}^t estimated in step 2, we can calculate the $\hat{\Lambda}^t(I_{tar},x,y)$ of Eq. (4) as:

$$\hat{\Lambda}(I_{tar}^t,x,y) = \int_{-\infty}^{M(I_{tar}^t|x,y)} \frac{1}{\sqrt{2\pi v_{out}^t}} e^{-\frac{(x-\mu_{out}^t)^2}{2v_{out}^t}} dx, \tag{11}$$

which is the probability of observing a confidence as low as the target updates under the null-hypothesis that the target point (x,y) is a non-member. Specifically, the small $\hat{\Lambda}(I_{tar}^t,x,y)$ indicates the target data is non-member.

Moreover, we utilize the updates of all communication rounds to obtain the $\tilde{\Lambda}(I_{tar}, x, y)$ of Eq. (5) as:

$$\begin{aligned}\tilde{\Lambda}(\{I^t_{tar}\}^T_{t=1}, x, y) &= \frac{1}{T}\sum_{t=1}^{T} \hat{\Lambda}(I^t_{tar}, x, y) \\ &= \frac{1}{T}\sum_{t=1}^{T} \int_{-\infty}^{M(I^t_{tar}|x,y)} \frac{1}{\sqrt{2\pi v^t_{out}}} e^{-\frac{(x-\mu^t_{out})^2}{2v^t_{out}}},\end{aligned} \tag{12}$$

Finally, given the threshold δ, the member is determined if $\tilde{\Lambda} > \delta$, otherwise is non-member.

Algorithm 1 FedMIA

Input: Target data sample (x, y) of client k, communication rounds T, set of client models $\{I^t_k | t \in [T], k \in [K]\}$, a threshold δ.
Output: membership prediction (0 or 1)
$\triangleright$ *Computing the low-dimensional measurement*
Calculate $M(I^t_k|x, y)$ according to Eq. (7) for all $k \in [K], t \in [T]$;
$\triangleright$ *Estimating the $\mathbb{Q}_{out}(I|x, y)$*
Choose $\mathcal{U}_t$ according to Eq. (8);
$\mu^t_{out} = \frac{1}{|\mathcal{U}^t|}\sum_{j\in\mathcal{U}^t} M(I^t_j|x, y)$;
$v^t_{out} = \frac{1}{|\mathcal{U}^t|}\sum_{j\in\mathcal{U}^t}(M(I^t_j|x, y) - \mu^t_{out})^2$;
$\triangleright$ *Inferring the membership*
Calculate $\hat{\Lambda}(I^t_{tar}, x, y)$ according to Eq. (11);
Calculate $\tilde{\Lambda}(\{I^t_{tar}\}^T_{t=1}, x, y)$ according to Eq. (12);
if $\tilde{\Lambda}(\{I^t_{tar}\}^T_{t=1}, x, y) > \delta$ **then**
    return 1
else
    return 0
end if

4. Experimental Result

This section presents the empirical analysis of the proposed FedMIA framework in terms of experimental setting, attack effectiveness, robustness.

4.1. Experimental Setup

Dataset & Models. We employed three image datasets: CIFAR-100 [19], DermNet [1], and Tiny-ImageNet [21]. Additionally, we implemented three models: ResNet18 [12] and AlexNet [20] for the classification task, and the Latent Diffusion Model [30] for the generative task.

Federated Setting. We consider horizontal federated learning with two typical structures: FedAvg [26], which transfers models or gradients in classification tasks, and FedEmbedding [24], which transfers prompts in generative tasks. We evaluate scenarios with 5–30 clients in FL, up to 300 communication rounds, and the data samples of each client range from 1,000 to 10,000. Local training involves 1 to 9 epochs, and we explore different Non-IID extents using the Dirichlet distribution $dir(\beta)$, with values of $\beta = 0.1, 1, 10, \infty$ (IID). If there are no additional instructions, each experiment has 10 clients, 300 synchronous communication rounds.

MIAs. We conducted a comprehensive comparison of our methods, FedMIA-I and FedMIA-II, against six baseline attack methods: Blackbox-Loss [40], Grad-Cosine [23], Grad-Norm [27], Loss-Series [10], Avg-Cosine [23], and Grad-Diff [23]. FedMIA-I is utilizes the model loss measurement [40], while FedMIA-II employs the Grad-Cosine measurement [23], as described in Eq. (7).

Defenses. We evaluate the robustness of FedMIA against six defense methods, including Gradient Perturbation (Perturb) [8, 45], Gradient Sparsification (Sparse) [11], MixUp [9, 43], Data Augmentation [34], Data Sampling [22], and a combination of Data Augmentation + Sampling.

Evaluation metric. We use the metrics AUC and TPR@FPR [3] to specifically assess the leakage of the most vulnerable samples to attacks, where TPR@FPR refers to the True Positive Rate (TPR) at a specific False Positive Rate (FPR, a.k.a. Type-I Error Rate) in binary classification. Specifically, we pay particular attention to the TPR when the FPR is very low, such as FPR values of 0.1%. Moreover, we test the attack effectiveness against different defense methods. Since the attack effectiveness depends on the parameters of defenses, e.g., for gradient perturbation, if more noise is added, the attack effectiveness becomes weaker, but the model utility is largely influenced. Therefore, we consider the attack effectiveness under different parameters of each defense method and obtain the tradeoff between utility loss (the test error rate) and attack effectiveness (TPR@FPR = 0.1%). Furthermore, the effectiveness of the attack can be measured using **hypervolume (HV)** [47], which, in our case, refers to the area between the Pareto frontiers and the unit box. A larger hypervolume indicates a better privacy-utility trade-off, representing that the attack is less effective.

More experimental setup details are left in Appendix A.

4.2. FedMIA vs Others

The comparison results of all attacks are presented in the Tab. 1. We can draw the following three conclusions:

- FedMIA generally outperforms other MIA methods across all experiments, as indicated by higher TPR@FPR=0.1% and AUC metrics. For example, FedMIA-II achieves a TPR@FPR=0.1% of (66.98 $\pm$ 1.74)%, which is significantly higher than the next best method with (54.66 $\pm$ 1.22)% on AlexNet-CIFAR100.

- Blackbox-Loss [40] and Grad-Cosine [23] show relatively low TPR and AUC values, significantly lagging

Table 1: Comparison of our attack with various MIAs methods on classification tasks and generative tasks. The larger TPR(%)@FPR=0.1% and AUC indicates the better attack effectiveness.

MIA Methods		Blackbox-Loss [40]	Grad-Cosine [23]	Grad-Norm [27]	Loss-Series [10]	Avg-Cosine [23]	Grad-Diff [23]	FedMIA-I (Ours)	FedMIA-II (Ours)
AlexNet CIFAR100	TPR	0.18±0.05	7.26±0.25	0.14±0.03	25.3±0.88	54.66±1.22	20.52±0.45	53.78±1.24	**66.98±1.74**
	AUC	0.58±0.01	0.78±0.02	0.51±0.01	0.82±0.01	0.85±0.01	0.61±0.02	**0.90±0.01**	0.89±0.02
AlexNet DermNet	TPR	0.27±0.12	9.53±0.54	0.13±0.03	22.8±1.13	41±0.48	5.6±0.05	48.23±0.87	**62.27±0.23**
	AUC	0.68±0.01	0.74±0.01	0.50±0.01	**0.94±0.01**	0.85±0.01	0.89±0.01	0.91±0.02	0.87±0.02
ResNet18 CIFAR100	TPR	0.36±0.11	5.48±0.27	0.26±0.06	16.82±2.12	44.02±1.58	15.06±1.78	57.36±2.12	**68.74±1.84**
	AUC	0.67±0.01	0.80±0.02	0.55±0.01	0.73±0.01	0.85±0.02	0.65±0.01	0.84±0.01	**0.89±0.01**
ResNet18 DermNet	TPR	0.27±0.11	0.73±0.21	0.06±0.01	32.2±0.89	19.93±1.23	17.93±2.12	**35.6±0.96**	31.8±0.88
	AUC	0.51±0.01	0.59±0.01	0.48±0.01	0.52±0.01	0.63±0.02	**0.66±0.01**	0.64±0.01	0.62±0.01
Diffusion Model Tiny-ImageNet	TPR	1.10±0.50	2.40±0.60	0.25±0.02	1.5±0.20	1.80±0.30	1.20±0.40	3.20±0.30	**4.50±0.20**
	AUC	0.51±0.02	0.54±0.01	0.49±0.01	0.54±0.01	0.53±0.01	0.51±0.01	0.58±0.01	**0.59±0.01**
Diffusion Model CIFAR100	TPR	0.80±0.10	1.20±0.20	0.11±0.01	1.30±0.10	1.80±0.10	1.7±0.20	2.1±0.20	**3.0±0.20**
	AUC	0.48±0.01	0.61±0.01	0.49±0.01	0.47±0.01	0.59±0.01	0.52±0.01	0.59±0.01	**0.62±0.01**

behind FedMIA. For example, in the AlexNet DermNet task, Blackbox-Loss and Grad-Cosine both result in low TPRs (around 0.27%), while FedMIA achieves a much higher TPR of (62.27 ± 0.23)%.

- FedMIA performs well in both classification tasks and generative task. Specifically, FedMIA also leads, but the performance gap between FedMIA and other methods like Blackbox-Loss or Grad-Cosine is more pronounced. For example, FedMIA achieves a TPR of (3.0 ± 0.20)% and an AUC of (0.62 ± 0.01)%, while Blackbox-Loss and Grad-Cosine have much lower TPR and AUC values (around (1.7 ± 0.20)% and (0.52 ± 0.01)%, respectively) in CIFAR100 with diffusion model.

(a) Original training images

(b) Generated images

Figure 3: Original training images and generated images based on uploaded embeddings via latent diffusion model.

Furthermore, we present the generated images based on the embeddings alongside the original images, which are considered as members by the proposed FedMIA-II, as shown in Fig. 3. The results indicate that the two types of images are highly similar, demonstrating that the target data identified by FedMIA as a member is effectively trained on the corresponding embedding. This observation highlights the effectiveness of the FedMIA attack.

4.3. Robustness

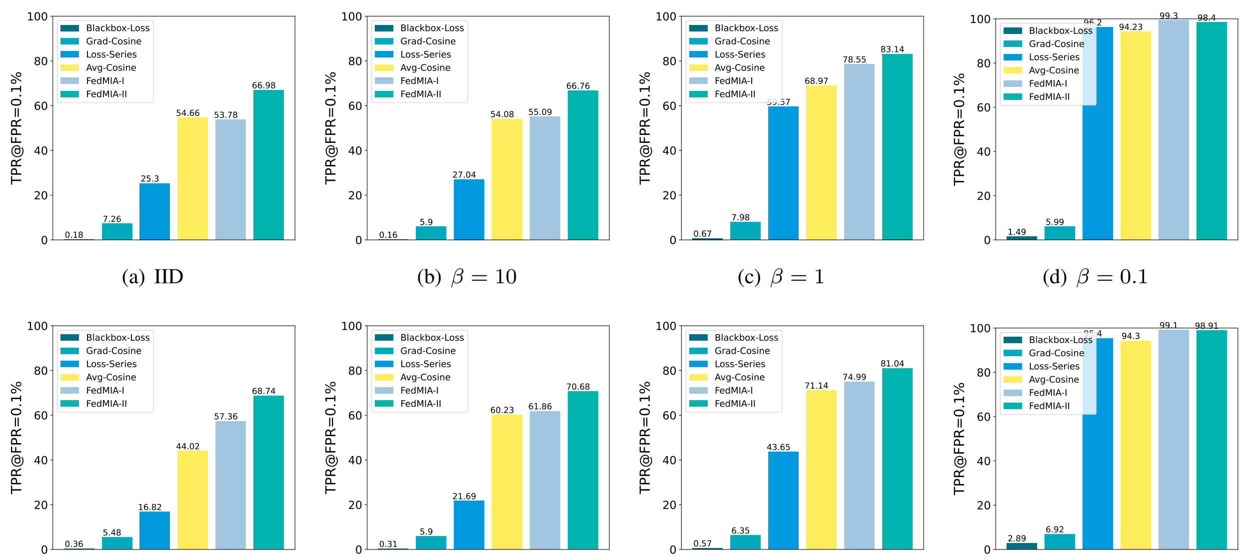
This section illustrates the robustness of FedMIA against six defense methods, varying degrees of Non-IID, different client counts, communication rounds, and local epochs.

FedMIA against different defense methods. Tab. 2 and Appendix B presents the hypervolume and privacy-utility tradeoff against different defense methods. We can obtain: 1)FedMIA-I (ours) consistently performs better than other MIA methods across all defense strategies for AlexNet: Example: For AlexNet-CIFAR100, under Mixup, FedMIA-I achieves a hypervolume of 0.3609, outperforming methods like Blackbox-Loss (0.3328) and Loss-Series (0.3554); 2) Combining data augmentation and data sampling combining in an appropriate manner may have the best defense effectiveness (achieves the largest HV volume); 3) Even the strongest defense with combining data augmentation and data sampling still exist privacy leakage when preserving the model performance (see Appendix B).

Non-IID extent. We investigate the impact of non-IID on MIA attacks. Following [14], the basic assumption of non-iid simulation in this part is that the labels of each client's training data follow the Dirichlet distribution. β is the core parameter controlling the distribution difference and the smaller the β, the greater the degree of non-iid. We report the performance of the FedMIA-II attack on the CIFAR-100 dataset in a table. We control the degree of non-IID by adjusting the parameter alpha, where a smaller β indicates a

Table 2: The hypervolume of various attack methods under defense strategies with AlexNet and ResNet on CIFAR100. The smaller Hypervolume indicates the better attack effectiveness.

		Perturb [45]	Sparse [11]	Mixup [43]	Sampling [22]	Data Aug [34]	Data Aug + Sampling
AlexNet CIFAR100	Blackbox-Loss [40]	0.3543	0.3533	0.3328	0.3491	0.3395	0.3422
	Loss-Series [10]	0.3252	0.3377	0.3554	0.3464	0.3375	0.3427
	Grad-Cosine [23]	0.304	0.3287	0.2845	0.3365	0.3247	0.3379
	Avg-Cosine [23]	0.3421	0.3421	0.3334	0.3407	0.3248	0.3376
	FedMIA-I (ours)	0.3611	0.3593	0.3609	0.3373	0.3202	**0.3359**
	FedMIA-II (ours)	**0.2702**	**0.3085**	**0.2588**	**0.3325**	**0.3175**	0.3361
ResNet CIFAR100	Blackbox-Loss [40]	0.4338	0.4307	0.4538	0.4568	0.5833	0.5815
	Loss-Series [10]	0.4126	0.4092	0.458	0.4555	0.5823	0.5821
	Grad-Cosine [23]	0.3261	0.3545	0.3905	0.4407	0.5751	0.5779
	Avg-Cosine [23]	0.4064	0.4059	0.4533	0.4537	0.571	0.5745
	FedMIA-I (ours)	0.4438	0.4392	0.4736	0.4446	**0.5665**	0.574
	FedMIA-II (ours)	**0.2969**	**0.3004**	**0.4302**	**0.4425**	0.5693	**0.5739**



(a) IID (b) $\beta = 10$ (c) $\beta = 1$ (d) $\beta = 0.1$

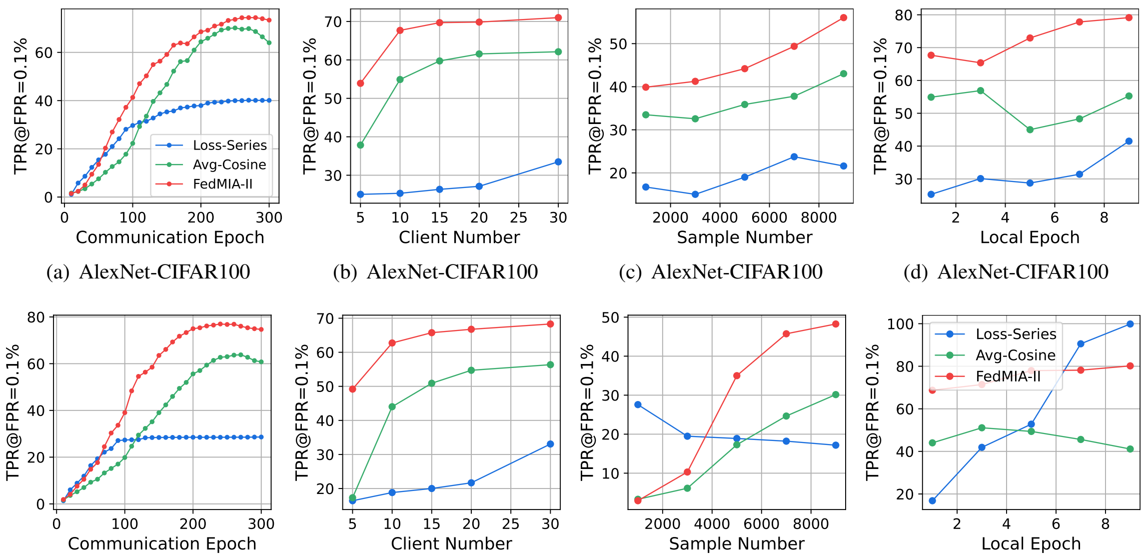
(e) IID (f) $\beta = 10$ (g) $\beta = 1$ (h) $\beta = 0.1$

Figure 4: This set of figures shows the attack effects (TPR@FPR=0.1%) of various attacks (Blackbox-Loss[40], Grad-Cosine[23], Loss-Series[10], Avg-Cosine[23], FedMIA-I and FedMIA-II) on AlexNet and ResNet18 (the first and second row respectively) under IID and three Non-IID settings.

more severe Non-IID condition.

Based on Fig.4, we can observe the following: 1) the proposed method is the strongest attacks with various Non-IID extent, e.g., the TPR@FPR =0.1% for FedMIA-II achieves the largest value 67% in AlexNet-CIFAR100. 2) As non-IID increases, TPR shows a increasing trend. One reason is when non-IID becomes severe, such as when a client contains only a few classes, MIA attacks themselves become easier. An extreme example is when a client contains only one class, in which case we can achieve a strong baseline by simply judging based on the sample labels.

Communication round. As for the benefits of synchronous rounds to our scheme, it can be observed in Figure 5(a) and (e) that the attack effect of our scheme increases rapidly in most epochs as the synchronous communication progresses. At epoch=200, the TPR@FPR=0.1% of Ours exceeds 0.4, which is twice as high as that of the Avg-Cosine attack and Loss-series attack. After that, the attack effect shows a slight decrease (about 5% on TPR@FPR=0.1%) and a similar trend can also be observed in the curve of the Avg-Cosine attack. This phenomenon may be attributed to the fact that the information obtained in the later epochs is not



(a) AlexNet-CIFAR100 (b) AlexNet-CIFAR100 (c) AlexNet-CIFAR100 (d) AlexNet-CIFAR100

(e) ResNet18-CIFAR100 (f) ResNet18-CIFAR100 (g) ResNet18-CIFAR100 (h) ResNet18-CIFAR100

Figure 5: This set of figures shows the attack effects (TPR@FPR=0.1%) of various attacks (blue line: Loss-series [10], green line: Avg-Cosine [23] and red line: FedMIA-II) on AlexNet and ResNet18 (the first and second row respectively) under four settings. The four columns of the graph group show the results of different communication rounds, client numbers, data volumes and local epochs settings respectively.

as helpful for the membership leakage attack as the information acquired in the previous epochs.

Number of Clients. Figure 5(b) and (f) illustrate the effects of membership inference attacks on AlexNet and ResNet18, while varying the number of clients from 2 to 20. As depicted in the figures, our two attacks outperform the baselines in most cases, indicating their significantly higher effectiveness. Furthermore, the gradually rising red and blue curves indicates that as the number of clients increases, the target model becomes more vulnerable to our MIA scheme.

Number of Samples. Figure 5(c) and (g) demonstrate the impact of varying the number of samples (ranging from 500 to 5000) on the attack effects of MIAs on AlexNet and ResNet18. Regardless of the increase in the number of samples, the attack effect of our scheme remains consistently high and even demonstrates notable improvement on AlexNet. The TPR@FPR=0.1% of ours consistently exceeds twice that of the baseline in both subfigures. This indicates that our attack scheme maintains a significant advantage as the training data increases.

Local Epoch. Figure 5(d) and (h) illustrate the effects of MIAs on AlexNet and ResNet18 as the number of local epochs varies from 1 to 9. As the number of local epochs increases, the effectiveness of our attack method and the fed loss attack significantly improve while the enhancement effect of the Avg-Cosine attack is not evident. This demonstrates that an increase in the number of local epochs may render the model more susceptible to MIA.

5. Discussion and Conclusion

While advantages brought by FL can be ascribed to, by and large, the principle of “one for all and all for one”, this paper shows that information shared by all clients through a semi-honest server can actually be adversely exploited to launch very effective membership inference attacks. Specifically, this paper introduce FedMIA, a novel Membership Inference Attack (MIA) method by leveraging updates from non-target clients and applies a one-tailed likelihood-ratio hypothesis test. This enables the inference of target data membership without requiring access to auxiliary datasets or making strong assumptions about the training process. Through extensive experiments, we demonstrated that FedMIA is highly effective across various federated learning configurations, including both classification and generative tasks, and remains robust against common defense methods, Non-IID data, and different client setups.

Conventional FL privacy defenses (perturbation, sparsification, mixup) prove ineffective against FedMIA due to attackers’ exploitation of cross-client information patterns as shown in Sect. 4.3. While secure aggregation via MPC [5][6]/HE [38][42] blocks FedMIA by encrypting individual updates, their computational/communication costs hinder practical deployment. This exposes an urgent need for MIA defenses specifically to resist attackers from obtaining valuable information from non-target updates.

References

[1] Amina Aboulmira, Hamid Hrimech, and Mohamed Lachgar. Comparative study of multiple cnn models for classification of 23 skin diseases. *International Journal of Online & Biomedical Engineering*, 18(11), 2022. 5

[2] Michael Aerni, Jie Zhang, and Florian Tramèr. Evaluations of machine learning privacy defenses are misleading. *arXiv preprint arXiv:2404.17399*, 2024. 2

[3] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022. 3, 5

[4] Christopher A Choquette Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. *arXiv preprint arXiv:2007.14321*, 2020. 2

[5] Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Helen Möllering, Thien Duc Nguyen, Phillip Rieger, Ahmad-Reza Sadeghi, Thomas Schneider, Hossein Yalame, et al. Safelearn: Secure aggregation for private federated learning. In *2021 IEEE Security and Privacy Workshops (SPW)*, pages 56–62. IEEE, 2021. 8

[6] Till Gehlhar, Felix Marx, Thomas Schneider, Ajith Suresh, Tobias Wehrle, and Hossein Yalame. Safefl: Mpc-friendly framework for private and robust federated learning. In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 69–76. IEEE, 2023. 8

[7] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020. 1

[8] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017. 5

[9] Hanlin Gu, Jiahuan Luo, Yan Kang, Lixin Fan, and Qiang Yang. Fedpass: Privacy-preserving vertical federated deep learning with adaptive obfuscation. *arXiv e-prints*, pages arXiv–2301, 2023. 5

[10] Yuhao Gu, Yuebin Bai, and Shubin Xu. Cs-mia: Membership inference attack based on prediction confidence series in federated learning. *Journal of Information Security and Applications*, 67:103201, 2022. 1, 4, 5, 6, 7, 8

[11] Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018. 5, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[13] Xinlong He, Yang Xu, Sicong Zhang, Weida Xu, and Jiale Yan. Enhance membership inference attacks in federated learning. *Computers & Security*, 136: 103535, 2024. 1

[14] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 6

[15] Hongsheng Hu, Xuyun Zhang, Zoran Salcic, Lichao Sun, Kim-Kwang Raymond Choo, and Gillian Dobbie. Source inference attacks: Beyond membership inference attacks in federated learning. *IEEE Transactions on Dependable and Secure Computing*, 2023. 2

[16] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. Practical blind membership inference attack via differential comparisons. *arXiv preprint arXiv:2101.01341*, 2021. 2

[17] Yan Kang, Hanlin Gu, Xingxing Tang, Yuanqin He, Yuzhu Zhang, Jinnan He, Yuxing Han, Lixin Fan, and Qiang Yang. Optimizing privacy, utility and efficiency in constrained multi-objective federated learning. *arXiv preprint arXiv:2305.00312*, 2023. 2

[18] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016. 1, 2

[19] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 5

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[21] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[22] Anran Li, Lan Zhang, Juntao Tan, Yaxuan Qin, Junhao Wang, and Xiang-Yang Li. Sample-level data selection for federated learning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021. 5, 7

[23] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Effective passive membership inference attacks in federated learning against overparameterized models. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 4, 5, 6, 7, 8

[24] Jinglin Liang, Jin Zhong, Hanlin Gu, Zhongqi Lu, Xingxing Tang, Gang Dai, Shuangping Huang, Lixin Fan, and Qiang Yang. Diffusion-driven data replay: A

novel approach to combat forgetting in federated class continual learning. In *European Conference on Computer Vision*, pages 303–319. Springer, 2025. 3, 5

[25] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017. 1, 2, 3

[26] H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016. 1, 2, 3, 5

[27] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018. 1, 4, 5, 6

[28] Milad Nasr, Reza Shokri, and Amir 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)*, pages 739–753. IEEE, 2019. 1, 2

[29] Shahbaz Rezaei and Xin Liu. Towards the infeasibility of membership inference on deep models. *arXiv preprint arXiv:2005.13702*, 2020. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[31] Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning (ICML)*. PMLR, 2019. 2

[32] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Annual Network and Distributed System Security Symposium (NDSS)*, 2019.

[33] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 2

[34] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 5, 7

[35] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. *arXiv preprint arXiv:2003.10595*, 2020. 2

[36] Anshuman Suri, Pallika Kanani, Virendra J Marathe, and Daniel W Peterson. Subject membership inference attacks in federated learning. *arXiv preprint arXiv:2206.03317*, 2022. 1

[37] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019. 2

[38] Febrianti Wibawa, Ferhat Ozgur Catak, Murat Kuzlu, Salih Sarp, and Umit Cali. Homomorphic encryption and federated learning based privacy-preserving cnn training: Covid-19 detection use-case. In *Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference*, pages 85–90, 2022. 8

[39] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 2, 3

[40] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018. 2, 5, 6, 7

[41] Oualid Zari, Chuan Xu, and Giovanni Neglia. Efficient passive membership inference attack in federated learning. *arXiv preprint arXiv:2111.00430*, 2021. 1, 2

[42] Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning. In *2020 USENIX annual technical conference (USENIX ATC 20)*, pages 493–506, 2020. 8

[43] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5, 7

[44] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. Gan enhanced membership inference: A passive local attack in federated learning. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2020. 1

[45] Qinqing Zheng, Shuxiao Chen, Qi Long, and Weijie Su. Federated f-differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2251–2259. PMLR, 2021. 5, 7

[46] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019. 1

[47] Eckart Zitzler and Simon Künzli. Indicator-based selection in multiobjective search. In *International con-*

ference on parallel problem solving from nature, pages 832–842. Springer, 2004. 5