

Membership Inference Attacks with False Discovery Rate Control

Chenxu Zhao Wei Qian Aobo Chen Mengdi Huai
Department of Computer Science
Iowa State University
{cxzhao, wqi, aobochoen, mdhuai}@iastate.edu

Abstract

Recent studies have shown that deep learning models are vulnerable to membership inference attacks (MIAs), which aim to infer whether a data record was used to train a target model or not. To analyze and study these vulnerabilities, various MIA methods have been proposed. Despite the significance and popularity of MIAs, existing works on MIAs are limited in providing guarantees on the false discovery rate (FDR), which refers to the expected proportion of false discoveries among the identified positive discoveries. However, it is very challenging to ensure the false discovery rate guarantees, because the underlying distribution is usually unknown, and the estimated non-member probabilities often exhibit interdependence. To tackle the above challenges, in this paper, we design a novel membership inference attack method, which can provide the guarantees on the false discovery rate. Additionally, we show that our method can also provide the marginal probability guarantee on labeling true non-member data as member data. Notably, our method can work as a wrapper that can be seamlessly integrated with existing MIA methods in a post-hoc manner, while also providing the FDR control. We perform the theoretical analysis for our method. Extensive experiments in various settings (e.g., the black-box setting and the lifelong learning setting) are also conducted to verify the desirable performance of our method.

1. Introduction

Deep neural networks (DNNs) have been successfully adopted in various computer vision tasks [18, 22, 25, 32, 40, 42, 68, 84, 85]. Due to the high sample complexity of such models, they require large amounts of training data. However, recent research highlights privacy risks and vulnerabilities of DNNs to membership inference attacks (MIAs) [37]. MIAs on DNNs aim to infer whether a specific data record was used to train a target model or not, thus posing severe privacy risks to individuals. For instance, if attackers infer that a clinical record has been used to train a model as-

sociated with a certain disease, they can infer that the owner of the record has that disease with high probability. A recent report [64] published by the National Institute of Standards and Technology (NIST) specifically highlights that an MIA revealing that an individual was included in the dataset used to train the target model is a confidentiality violation.

On the other hand, beyond traditional privacy attacks, MIAs have diverse applications and play a crucial role in fields such as machine unlearning [17, 65] and lifelong learning [71]. For example, in the context of machine unlearning, MIA methods are used to evaluate the effectiveness of unlearning methods and help verify that specific samples have been successfully unlearned, thus respecting individuals' rights to have their data removed. Note that machine unlearning refers to the process of selectively removing the influence of specific samples from trained models [11, 15, 29, 53, 72, 83]. Additionally, in lifelong learning, where a model aims to learn continuously from new data over time while retaining previously acquired knowledge, MIAs can be used to gauge the degree of memorization for certain data. This allows us to assess how well the model retains specific information and preserves prior knowledge.

Currently, various MIA methods have been proposed. Based on the differentiation principles, existing MIA methods can be generally divided into: *classifier-based*, *metric-based*, *likelihood ratio-based*, and *quantile regression-based*. Specifically, classifier-based MIAs [16, 34, 61] usually train a binary membership classifier indicator to distinguish the behavior of training members from that of non-training members. Metric-based MIAs [20, 23, 41] involve defining a specific metric on model outputs to distinguish training members from non-members. Likelihood ratio-based MIAs [12, 78] utilize parametric techniques to model the loss distributions of models that have been trained or not trained on the target test example. Quantile regression-based MIAs [10] utilize quantile regression on non-member distributions without training surrogate models.

Despite the significance and popularity of MIAs, existing MIA methods cannot provide guarantees on the false discovery rate (FDR), which is defined as the expected

proportion of instances classified as training data (members) but are, in reality, not part of the training data (non-members) among total instances classified as training data. Traditional MIA works usually focus on empirical comparisons, and fail to provide the theoretical guarantees for member and non-member decisions. Although [10, 75] consider the ratio of non-members incorrectly identified as members, they cannot provide the guarantees on the false discovery rate. In practice, the false discovery rate provides a more precise indication of the error rate [5, 8, 9, 39], and is crucial in settings involving simultaneous evaluations. By managing the FDR, we can mitigate the risks associated with evaluating the reliability of positive discoveries and make more informed decisions under conditions of uncertainty [7, 27, 47, 49, 50]. Additionally, existing MIAs also fail to provide the marginal probability guarantee on labeling true non-member data as member data.

Our goal in this paper is to provide the guarantees on the false discovery rate for MIAs, which refers to the proportion of false discoveries among total positive discoveries in the overall testing procedure. Note that existing works on MIAs are typically framed as a hypothesis testing problem, with the alternative hypothesis asserting that the test data is from the training dataset and the null hypothesis positing it is not. However, managing the false discovery rate in the context of MIAs presents unique challenges. First, the distribution of scores for non-training data remains unknown and challenging to model accurately, which complicates the derivation of the membership indicators. Additionally, the estimated non-member probabilities usually exhibit interdependence. However, traditional multiple hypothesis testing techniques usually assume that inputs must either be independent or adhere to certain conditions of dependency. This assumption is hard to be satisfied in practice, which complicates the process of accurately controlling the FDR and makes it difficult to ensure that the proportion of false discoveries remains within acceptable limits.

To address the above challenges, in this paper, we propose *MIAFDR*, a novel membership inference attack that can provide the false discovery rate guarantees. Specifically, in our method, given that the underlying true distribution of the member data and that of the non-member data are hard to know, we first design a novel conformity score function, which can reflect the conformity degree of test data to the non-member data. Next, based on estimated point-wise conformity scores, we present a non-member relative probability estimation strategy, which essentially reflects the likelihood of not making discoveries. We also show that based on these estimated non-member probabilities, we can provide the marginal probability guarantee on labeling true non-member data as member data. However, these generated point-wise non-member relative probabilities exhibit interdependence, making it challenging to provide the false

discovery rate control. To address this, we then present an adjustment method that corrects these calculated non-member probabilities by accounting for their interdependencies and employing a weighted correction scheme. We also show that these adjusted non-member probabilities allow for controlling the false discovery rate at a predetermined significance level for prediction results. Notably, our method can be seamlessly integrated with existing MIAs to provide FDR control while preserving their attack performance. We conduct the theoretical analysis for our method. Our extensive experiments verify the effectiveness of our method. We also empirically show that our method can help data memorization-based machine learning (ML) tasks, including machine unlearning and lifelong learning.

2. Related Work

Membership inference attacks (MIAs) are designed to determine whether a given data sample has been used to train a particular model. The concept of MIAs is first proposed by [36], which aims to detect sensitive and private information leakage. Specifically, this work trains multiple shadow models to mimic the behavior of the victim model to distinguish between training samples from the training dataset and test samples. Since their inception, MIAs have gained significant attention in the research community, leading to the development of numerous MIA methods [10, 12, 20, 23, 34, 41, 48, 55, 81]. Notably, numerous studies have sought to elucidate mechanisms behind MIAs, primarily attributing their operations to model memorization [2, 13, 26, 31, 55, 59], a phenomenon linked to overfitting. Due to memorization in DNNs, prediction confidence tends to be higher for data used for training. This difference in prediction confidence helps MIA methods to determine which image data were used for training. In [76], the authors theoretically analyze the relationship between overfitting and MIAs. Therefore, beyond detecting sensitive information leakage, MIAs can also offer valuable insights into the extent of memorization in the victim model.

Currently, MIAs have been successfully achieved in many domains and problems, including semantic image segmentation [35, 79], healthcare [30, 74], image classification [37, 55], and recommendation systems [19, 80]. For example, [35] shows that such membership inference attacks can be successfully carried out on state-of-the-art models for semantic segmentation. However, existing MIAs cannot guarantee the false discovery rate, which refers to the proportion of false discoveries among total discoveries during the attack procedure. This oversight is particularly problematic when the proportion of actual members within the test data is high. In this work, we build upon conformal inference [6, 7, 14, 28, 44, 46, 51, 54, 66, 67, 69, 77], which aims to quantify the uncertainty in predictions with a specified coverage probability. However, traditional conformal infer-

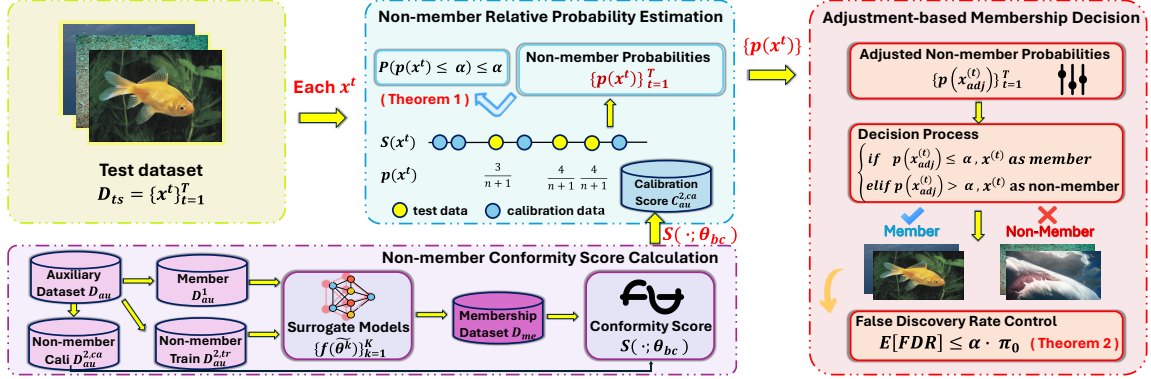


Figure 1. Overview of our proposed membership inference attacks with the false discovery rate guarantee.

ence works cannot be directly applied to provide the false discovery rate guarantees for MIAs. As aforementioned, this limitation arises because calculated non-member probabilities for test data are conditional on the shared calibration dataset and exhibit interdependence, making it challenging to provide the false discovery rate control. In contrast, our method can provide the false discovery rate guarantee, even under existing defenses against MIAs [1, 52, 60].

3. Threat Model

We consider a threat model that includes a model holder and an adversary. The model holder owns a well-trained DNN classifier $f(\theta^*)$, which is trained over its training data $D_{tr} = \{(x_i, y_i)\}_{i=1}^N$ using the loss function \mathcal{L} (e.g., cross-entropy loss), where $x_i \in \mathbb{R}^d$ is a d -dimensional feature vector and $y_i \in [Y]$ denotes its associated label. For a given input x , we can generate its softmax probability vector as

$$f(x; \theta^*) = [f_1(x; \theta^*), f_2(x; \theta^*), \dots, f_Y(x; \theta^*)]. \quad (1)$$

For the given input data x , we can determine its class label as $y(x) = \arg \max_{y \in [Y]} f_y(x; \theta^*)$, where a label $y \in [Y]$ with the largest probability is assigned to this given data x .

The adversary aims to distinguish the training (members) and test data (non-members) of the victim model. Let $D_{ts} = \{x^t\}_{t=1}^T$ denote the test dataset available to the adversary. As shown in Definition 1, for each $x^t \in D_{ts}$, the adversary's goal of determining whether $x^t \in D_{tr}$ or $x^t \notin D_{tr}$ can be formulated as a Hypothesis Testing, where the null hypothesis (i.e., H_0) represents non-membership, while the alternative hypothesis H_1 represents membership.

Definition 1 (Hypothesis Testing for MIA). *Let x^t denote the test data. Let θ^* denote the pre-trained model trained on its private training dataset D_{tr} . Then we can establish a hypothesis test where the null hypothesis posits that x^t was not part of the private training dataset D_{tr} for the target victim model θ^* , as outlined below*

$$H_0 : x^t \notin D_{tr}, \quad H_1 : x^t \in D_{tr}, \quad (2)$$

where D_{tr} is the private training dataset for $f(\theta^*)$.

For the adversary's knowledge, we consider two realistic attack settings: *grey-box* and *black-box*.

- In the grey-box setting, we assume that the adversary does not have access to the exact private training dataset D_{tr} , and has no knowledge of the victim model parameters. We consider that the adversary is aware of the knowledge of the owner's learning algorithm and architecture.
- In the black-box setting, we assume that the adversary does not have any prior knowledge about the private training dataset, the learning algorithm, or the target pre-trained model, including its architecture and parameters. This setting produces a realistic threat model in real-world applications, where adversaries typically operate with minimal information.

In the above two settings, we consider a realistic adversary who has limited knowledge and does not have access to the private training dataset D_{tr} . However, we allow the adversary to have access to an auxiliary dataset D_{au} , distinct from the private dataset, sampled from the same distribution. This assumption is reasonable given the widespread availability of public data, and has been a common assumption for black-box attacks in existing literature [38]. Notably, our proposed MIAFdr builds upon existing MIA methods, serving as a wrapper to ensure FDR control. Since existing MIAs [10, 12, 34, 61, 82] typically leverage on auxiliary data for training surrogate models or regression models, we similarly integrate auxiliary data into MIAFdr, aligning with the established practices. By exploring these two settings, we aim to gain comprehensive insights into the different levels of threats posed by attackers with varying degrees of knowledge about the target model.

4. Methodology

In this section, we utilize the classifier-based setting to present our method. Note that our method serves as a wrapper for existing MIAs with FDR control. *Discussions on other settings are deferred to the full version of the paper.*

Overview. Figure 1 shows the overall framework of our proposed method, which can ensure control over the false

discovery rate. Specifically, our method involves three essential components: *non-member conformity score calculation*, where we design a novel conformity score function to assess the degree to which each test sample conforms to the non-member data distribution; *non-member relative probability estimation*, where we utilize the previously calculated non-member conformity score to estimate the non-member relative probability for each test data, which can reflect the likelihood of this given test data being non-training data; and *adjustment-based membership decision*, where we adjust the previously estimated related probabilities for these test data samples and then compare them against a pre-defined significance level to make membership decisions. Below, we will detail each of the three essential components.

Non-member Conformity Score Calculation. Note that based on Definition 1, to determine whether x^t comes from D_{tr} , we can establish the following hypothesis test: the null hypothesis (i.e., H_0) posits that x^t is not from the private training dataset D_{tr} (non-members) for the victim model $f(\theta^*)$, and the alternative hypothesis (i.e., H_1) posits that it is from the private dataset. By leveraging the information of the target victim model $f(\theta^*)$, we can reduce this hypothesis testing to the problem of determining whether the output (i.e., $f(x^t; \theta^*)$) of the test data x^t belongs to the final softmax output distribution of the victim model θ^* . However, in practice, it is very difficult to obtain its underlying true final output distribution. To address this, we will estimate the empirical output distribution by considering the private training samples' final output predictions (i.e., $\{f(x_i; \theta^*) : x_i \in D_{tr}\}_{i=1}^N$). Then we can estimate the empirical probability distribution $\hat{f}_N(\nu)$ via the average of delta functions, where ν is any real number. Thus, we can construct the below null hypothesis testing

$$\tilde{H}_0 : f(x^t; \theta^*) \not\sim \hat{f}_N(\nu) = \frac{1}{N} \sum_{i=1}^N \delta(\nu - f(x_i; \theta^*)), \quad (3)$$

where δ is the Dirac delta function, which returns ∞ if the condition $f(x_i; \theta^*) = \nu$ is true, and 0 otherwise.

However, in practice, the attacker usually does not have access to the private training dataset D_{tr} for the victim model $f(\theta^*)$, which makes it difficult to characterize the population of non-members. Therefore, it is intractable to directly adopt the hypothesis testing constructed in Eqn. (3). To address this, we first train K surrogate models (denoted as $\{f(\tilde{\theta}^k)\}_{k=1}^K$) and collect their predictions to estimate the empirical score distribution of non-members. Specifically, as shown in Figure 1, to obtain K surrogate models, we first divide the auxiliary dataset D_{au} into two disjoint subsets, i.e., D_{au}^1 and D_{au}^2 , where $D_{au}^1 \cap D_{au}^2 = \emptyset$ and $D_{au}^1 \cup D_{au}^2 = D_{au}$. From D_{au}^1 , we will create K subsets (i.e., $\{D_{au}^{1,k}\}_{k=1}^K$) by sampling a fraction η of the data without replacement. Then, we can optimize the k -th surrogate model as $\tilde{\theta}^k \leftarrow \arg \min_{\theta} \sum_{(x_i, y_i) \in D_{au}^{1,k}} \mathcal{L}((x_i, y_i); \theta)$,

where \mathcal{L} is the victim model's loss assumed in the grey-box setting. Thus, we can train K surrogate models (i.e., $\{f(\tilde{\theta}^k)\}_{k=1}^K$), which can approximate the behavior of the victim model $f(\theta^*)$.

Based on these K surrogate models (i.e., $\{f(\tilde{\theta}^k)\}_{k=1}^K$), for all samples within $D_{au}^{1,k}$, we can obtain their predictions $\mathcal{Y}_{au}^{1,k} = \{f(x_i; \tilde{\theta}^k)\}_{i=1}^{|D_{au}^{1,k}|}$, where $\tilde{\theta}^k$ is the k -th surrogate model trained on $D_{au}^{1,k}$. We then split the D_{au}^2 into two disjoint sets $D_{au}^{2,tr}$ and $D_{au}^{2,ca}$, where $D_{au}^{2,tr} \cap D_{au}^{2,ca} = \emptyset$, and $D_{au}^{2,tr} \cup D_{au}^{2,ca} = D_{au}^2$. Similarly, for each sample $x_i \in D_{au}^{2,tr}$ and $x_j \in D_{au}^{2,ca}$, we can obtain $\mathcal{Y}_{au}^{2,i} = \{f(x_i; \tilde{\theta}^k)\}_{k=1}^K$ and $\mathcal{Y}_{au}^{2,j} = \{f(x_j; \tilde{\theta}^k)\}_{k=1}^K$. Thus, we have

$$\mathcal{Y}_{au}^1 = \cup_{k=1}^K \mathcal{Y}_{au}^{1,k}, \quad \mathcal{Y}_{au}^{2,tr} = \cup_{i=1}^{|D_{au}^{2,tr}|} \mathcal{Y}_{au}^{2,i}, \quad (4)$$

$$\text{and } \mathcal{Y}_{au}^{2,ca} = \cup_{j=1}^{|D_{au}^{2,ca}|} \mathcal{Y}_{au}^{2,j},$$

where $D_{au}^{2,tr} \cup D_{au}^{2,ca} = D_{au}^2 \subset D_{au}$. Based on this, we will construct the below membership dataset

$$D_{me} = \{(y_i, 0) : y_i \in \mathcal{Y}_{au}^1\} \cup \{(y_i, +1) : y_i \in \mathcal{Y}_{au}^{2,tr}\}.$$

The above constructed membership dataset D_{me} can effectively capture the member distribution using the member samples labeled as 0 and the non-member prediction using the non-member samples labeled as +1.

Based on the constructed membership dataset $D_{me} = \{z_i = (y_i, l_i)\}_{i=1}^{|\mathcal{Y}_{au}^1| + |\mathcal{Y}_{au}^{2,tr}|}$, where $l_i \in \{0, +1\}$, we will train the following binary classifier $f_{bc}(\theta_{bc})$ to distinguish between members and non-members

$$\theta_{bc} = \arg \max_{\theta} \sum_{z_i \in D_{me}} \mathcal{L}_{bc}(z_i = (y_i, l_i); \theta), \quad (5)$$

where \mathcal{L}_{bc} is the loss for training this classifier. Note that for $z_i = (y_i, l_i) \in D_{me}$, it is either labeled $l_i = 0$ (members) or $l_i = +1$ (non-members). Then, for the given test data x^t , to reflect how typical it is with respect to the calibration data, we define the below non-member conformity score function

$$S(y^t; \theta_{bc}) = \lambda \log\left(\frac{f_{bc}(y^t; \theta_{bc})}{1 - f_{bc}(y^t; \theta_{bc})}\right) + (1 - \lambda) f_{bc}(y^t; \theta_{bc}), \quad (6)$$

where $y^t = f(x^t; \theta^*)$, $f_{bc}(\theta_{bc})$ denotes the trained binary classifier based on Eqn. (5) and λ is a hyper-parameter to control the weight between the logit-transformed probability and raw probability. Note that for the test data x^t , a larger non-member conformity score $S(y^t; \theta_{bc}) \in \mathbb{R}$ means that it is coming from the non-member prediction; otherwise, it is more likely to be from the member distribution.

Non-member Relative Probability Estimation. Based on the above conformity score function $S(\cdot; \theta_{bc})$, for all the samples within the set $\mathcal{Y}_{au}^{2,ca}$, we can calculate their conformity scores as $\mathcal{C}_{au}^{2,ca} = \{S(y_i; \theta_{bc}) : y_i \in \mathcal{Y}_{au}^{2,ca}\}_{i=1}^{|\mathcal{Y}_{au}^{2,ca}|}$. Then, we define P_{nm}^* as the true distribution of these conformity scores $\mathcal{C}_{au}^{2,ca} = \{S(y_i; \theta_{bc})\}_{i=1}^{|\mathcal{Y}_{au}^{2,ca}|}$. To determine

whether the given test data x^t is from the private dataset D_{tr} , we reformulate the hypothesis test in Eqn. (3) into

$$\hat{H}_0 : S(y^t; \theta_{bc}) \sim P_{nm}^*, \hat{H}_1 : S(y^t; \theta_{bc}) \not\sim P_{nm}^*, \quad (7)$$

where P_{nm}^* is the underlying true distribution of these conformity scores $\mathcal{C}_{au}^{2,ca} = \{S(y_i; \theta_{bc}) : y_i \in \mathcal{Y}_{au}^{2,ca}\}_{i=1}^{|\mathcal{Y}_{au}^{2,ca}|}$.

However, the underlying true distribution P_{nm}^* is usually unknown, which presents significant challenges for performing the hypothesis testing procedure in Eqn. (7). On the other hand, directly adopting existing empirical distribution estimation methods cannot provide theoretical guarantees for the false discovery rate control, and they also usually require the assumption of underlying distributions. To address this, as demonstrated in Figure 1, instead of estimating underlying distributions, we propose to calculate the below non-member relative probability for the test data x^t

$$p(x^t) = \frac{|\mathbb{S}^k \in \mathcal{C}_{au}^{2,ca} \cup \{S(y^t; \theta_{bc})\} : \mathbb{S}^k \leq S(y^t; \theta_{bc})|}{1 + |\mathcal{C}_{au}^{2,ca}|}, \quad (8)$$

where $\mathcal{C}_{au}^{2,ca} = \{S(y_i; \theta_{bc})\}_{i=1}^{|\mathcal{Y}_{au}^{2,ca}|}$, $y^t = f(x^t; \theta^*)$, and $S(y^t; \theta_{bc})$ is the conformity score for the test data x^t . The calculated non-member relative probability $p(x^t)$ is essentially the proportion of the calibration samples with conformity scores smaller than or equal to that of x^t . This can reflect the conformity degree of x^t to the non-member calibration scores. Thus, we can use the calculated non-member probability $p(x^t)$ to assess the likelihood that x^t was not used to train the target victim model $f(\theta^*)$.

Theorem 1. *Let \mathcal{G} denote the sequence containing all the samples from dataset $D_{au}^{2,ca}$ and the given test data x^t , i.e., $\mathcal{G} = (x^1, x^2, \dots, x^{|\mathcal{D}_{au}^{2,ca}|}, x^{|\mathcal{D}_{au}^{2,ca}|+1})$, where $|\mathcal{D}_{au}^{2,ca}|$ is the number of samples in $D_{au}^{2,ca}$, and $x^{|\mathcal{D}_{au}^{2,ca}|+1}$ is the extra term represented by x^t . Note that $D_{au}^{2,ca}$ is a subset of the auxiliary dataset D_{au} . Assume that this sequence is exchangeable. Then, for significance level $\alpha \in (0, 1)$, we have*

$$\mathcal{P}(p(x^t) \leq \alpha \mid x^t \notin D_{tr}) \leq \alpha, \quad (9)$$

where $p(x^t)$ is the calculated non-member probability for test data x^t , and D_{tr} represents the private training dataset.

In Theorem 1, we show that for a true non-member test data x^t , the probability that its non-member relative probability $p(x^t)$ is not larger than α is at most α . For this true non-member data x^t , it does not come from the private dataset D_{tr} (i.e., $x^t \notin D_{tr}$). This ensures that the error rate for labeling true non-member data as member data does not exceed the predefined threshold α , providing a guarantee on the reliability of the membership labeling process. Note that the exchangeability assumption in Theorem 1 is much less restrictive than the traditional independent and identically distributed (i.i.d.) assumption [4, 7, 21]. The proof for Theorem 1, and more discussions of our method on other MIA settings are deferred to the full version of the paper.

Adjustment-based Membership Decision. Next, we discuss how to provide the false discovery rate control for the test dataset $D_{ts} = \{x^t\}_{t=1}^T$, based on the above estimated non-member relative probabilities. From Definition 2, we can see that the false discovery rate is the expected proportion of false discoveries among all discoveries. However, these calculated p-values (i.e., non-member probabilities $\{p(x^t)\}_{t=1}^T$) for the test data D_{ts} are conditional on the calibration dataset $D_{au}^{2,ca}$; specifically, it applies uniformly across all test data, as each calculation involves the identical calibration dataset. Consequently, the p-values obtained for the membership inference attack exhibit interdependence, which makes it challenging to provide the false discovery rate control. To control the number of false discoveries among total discoveries in the overall testing procedure, we will adjust the calculation of the original p-value in Eqn. (8). Specifically, we first arrange these calculated $\{p(x^t)\}_{t=1}^T$ in ascending order, and obtain the ranked set $\{p^{(t)}\}_{t=1}^T$, where $p^{(t)}$ is the non-member probability ranked at position t . Subsequently, for $p^{(t)}$, as illustrated in Figure 1, we calculate its adjusted non-member probability as

$$p_{\text{adj}}^{(t)} = \min\{1, \min_{m \in \{t, t+1, \dots, n\}} \{\frac{n}{m} \cdot p^{(m)}\}\}, \quad (10)$$

where $p^{(m)} \in \{p^{(t)}\}_{t=1}^T$ ranks at position m .

Definition 2 (False Discovery Rate (FDR)). *Let A denote the MIA attack algorithm, where $A(x^t) \in \{0, 1\}$, with 0 indicating that x^t is a member of the private training dataset D_{tr} of the target victim model and 1 indicating it is not. Let D_{ts} denote the test dataset. Then we can define the false discovery rate as $\xi = |v_{fp}| / (|v_{fp}| + |v_{tp}|)$, where $v_{fp} = \{x^t \in D_{ts} \mid A(x^t) = 0, \text{ and } x^t \notin D_{tr}\}$ and $v_{tp} = \{x^t \in D_{ts} \mid A(x^t) = 1, \text{ and } x^t \in D_{tr}\}$.*

Based on these adjusted probabilities, for each test data $x^{(t)} \in D_{ts}$, we can obtain its null hypothesis $H_{0,(t)}$, i.e., $H_{0,(t)} : x^{(t)} \notin D_{tr}$. Then, for $x^{(t)}$, we can determine

$$\begin{cases} \text{if } p_{\text{adj}}^{(t)} \leq \alpha, & H_{0,(t)} \text{ does not hold,} \\ \text{if } p_{\text{adj}}^{(t)} > \alpha, & H_{0,(t)} \text{ holds,} \end{cases} \quad (11)$$

where $p_{\text{adj}}^{(t)}$ is the adjusted non-member probability calculated by Eqn. (10) and α is a pre-defined value for determining the significance level for the multiple hypothesis testing. In the above equation, if $p_{\text{adj}}^{(t)} \leq \alpha$, we should reject the null hypothesis, suggesting that the test data $x^{(t)}$ does not belong to the non-members and comes from the training data D_{tr} of the victim model θ^* ; otherwise, we should accept the null hypothesis, indicating that $x^{(t)} \notin D_{tr}$. Based on the above, for test dataset D_{ts} , we can obtain following decisions

$$\mathcal{R}(D_{ts}) = \{t : t \in [T], \text{ and } p_{\text{adj}}^{(t)} \leq \alpha\}, \quad (12)$$

where $p_{\text{adj}}^{(t)}$ is the adjusted non-member probability for the sample ranked at position (t) and α is the significance level.

Note that $\mathcal{R}(D_{ts})$ is the set of indices of test data within the test dataset D_{ts} for which the null hypothesis is rejected. This indicates that these test samples are likely members of the private training dataset D_{tr} .

Theorem 2. Let $\mathcal{H}_0^*(D_{ts}) = \{t : t \in [T], \text{ and } H_{0,(t)}^* \text{ is true}\}$ denote the subset of true non-members in the test data D_{ts} , where $H_{0,(t)}^*$ is the ground truth. Let π_0 denote the proportion of true non-members, i.e., $\pi_0 = \mathcal{H}_0^*(D_{ts})/T$. Then we can control the false discovery rate (FDR) at level $\alpha \cdot \mathcal{H}_0^*(D_{ts})/T$ as follows

$$\mathbb{E}\left[\frac{|\mathcal{R}(D_{ts}) \cap \mathcal{H}_0^*(D_{ts})|}{\max\{1, |\mathcal{R}(D_{ts})|\}}\right] \leq \alpha \cdot \frac{\mathcal{H}_0^*(D_{ts})}{T} \leq \alpha, \quad (13)$$

where $\mathcal{R}(D_{ts})$ is the obtained positive discovery results, and α is the significance level in Eqn. (12).

Theorem 2 states that our method allows for controlling the false discovery rate at a predetermined level for the generated positive discovery results (i.e., $\mathcal{R}(D_{ts})$ for test dataset D_{ts}), thereby limiting the proportion of false discoveries among the total discoveries [7]. The proof for Theorem 2, and more discussions of our method on other MIA settings are deferred to the full version of the paper.

Discussion. Note that in the above, we focus on the grey-box setting. For the threat of MIAs, we also consider the black-box setting, where attackers have no prior knowledge of the target model, including its private training data and model architecture. In this black-box setting, attackers can utilize varying architectures to train on the auxiliary dataset to approximate the target model’s behavior and obtain score distributions. This strategy leverages the transferability property that arises from shared decision boundaries across different models [45, 58, 70]. In this way, attackers can effectively conduct MIAs in the black-box scenario.

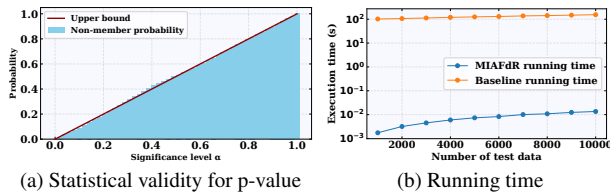


Figure 2. Statistical validity and running time.

5. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed MIAFdr. *More experimental details and results (e.g., Quantile Regression-based MIAs and differential privacy-based MIA defenses) are deferred to the full version of the paper.*

Datasets and Models. In the experiments, we adopt the following popular benchmark image datasets: Tiny-ImageNet [24], CIFAR-100 [43], and CIFAR-10 [43]. Additionally, our experimental evaluations are also conducted

on various deep learning models, including ResNet-50, ResNet-18 [33], VGG-16 [62], MobileNetV2 [57], and a multi-layer perceptron (MLP) network.

Baselines. In experiments, we compare our method with the following popular MIAs: classifier-based method such as shadow training [61]; metric-based methods, including Softmax [56], Modified Entropy (Entropy) [63], Loss [76], and Difficulty Calibration (Calibration) [73]; likelihood ratio-based method represented by Likelihood Ratio Attack (LiRA) [12]; and quantile regression-based method like Quantile Regression Attack [10].

Attack Setup. In experiments, we split the available dataset into: a private set used for training a target model, accessible only to the model holder, and a public set employed for training a surrogate model and querying conformity scores for attackers. In likelihood ratio-based and quantile regression-based methods, the public set serves as an auxiliary set for computing conformity scores. In classifier-based and metric-based methods, we allocate 40% of data for calibration and 60% for training the discriminative model using MLP. The evaluation is repeated 10 times, with the mean and standard deviation reported.

5.1. Attack Performance

First, we perform experiments to validate the Theorem 1, which establishes the validity of p-values generated by Eqn. (8). We employ our proposed method with the shadow training technique [61] from classifier-based MIAs, utilizing the ResNet-18 model on the CIFAR-10 dataset. Initially, we partition 30% of the auxiliary dataset D_{au} to form D_{au}^1 , allocating the remaining 70% to D_{au}^2 . From D_{au}^2 , we then sample a fraction $\eta = 3/7$ to obtain $\{D_{au}^{1,k}\}$. For each test data, we classify it as a member if its non-member probability is less than or equal to a predefined threshold α . As shown in Figure 2a, our approach ensures that the expected error rate for false discoveries among non-training data does not exceed α , thereby guaranteeing the reliability of the membership labeling process.

In addition, we aim to demonstrate the computation efficiency of our proposed method and the effectiveness of the adjusted non-member probabilities in managing the final membership decisions with respect to FDR. We follow the same experimental setup in Figure 2a. Notably, our MIAFdr serves as a wrapper for existing MIAs and maintains the same training time. In Figure 2b, we present an analysis of the inference runtime for the baseline method and our approach. The experimental results in Figure 2b show that our approach requires only a minimal additional running time compared to the baseline. For instance, with 7,000 test samples, the traditional classifier-based MIA takes 137.84 seconds, while our method adds only 0.01 seconds to the overall running time. In Figure 3, we present the results of the FDR against varying significance level α for $\pi_0 = 0.25$ and $\pi_0 = 0.5$ in both classifier-based and metric-based (Soft-

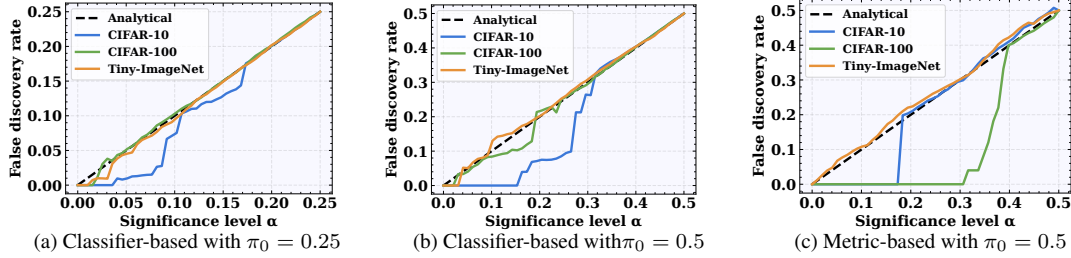


Figure 3. FDR control of classifier-based and metric-based MIAFDR.

max) settings, where π_0 is the proportion of non-members in the test dataset. From this figure, we can see that the FDR is mostly bounded by α , indicating that our method can effectively produce membership decisions with FDR control.

Next, we examine the effectiveness of our method in the presence of existing MIA defenses. We first evaluate the MIAFDR performance against the knowledge distillation-based (KD) defense. Specifically, we first train a teacher model and use its soft outputs with temperature $T = 20$ alongside the hard labels to train the student via a combined cross-entropy and KL divergence objective. The results are presented in Figure 4. From Figure 4a, we can see that our method successfully maintains control over FDR even under MIA defenses. This underscores the validity of our estimated non-member relative probability and the adjustment-based membership decision, confirming their robustness even under defenses. Additionally, Figure 4b illustrates that our method can preserve MIA prediction accuracy and AUROC under MIA defenses. More experimental results on MIAFDR against existing differential privacy-based defenses can be found in the full version of the paper.

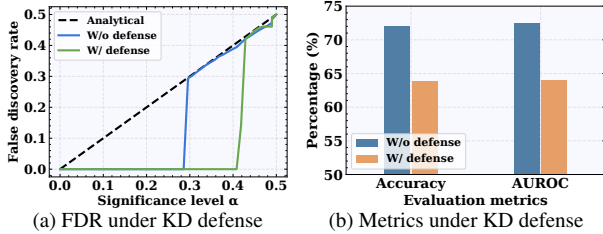


Figure 4. MIAFDR against KD defense.

Moreover, we investigate the attack effectiveness of MIAFDR with classifier-based MIAs, extending our analysis beyond the FDR control. Our assessment contrasts the baseline classifier-based MIA [61] with an enhanced version that incorporates our proposed MIAFDR across multiple datasets. The experimental results, presented in Table 1, further demonstrate that our MIAFDR achieves comparable or even superior attack performance than the baseline classifier-based MIA in terms of attack accuracy and the AUROC score. For instance, on the CIFAR-100 dataset, our MIAFDR method attains an attack accuracy of approximately 78.2%, outperforming the baseline accuracy

of 76.8%. These outcomes underscore the effectiveness of our classifier-based MIAFDR in accurately identifying data membership across member and non-member distributions.

Dataset	Method	Accuracy (%)	AUROC (%)
CIFAR-100	Classifier	76.81 \pm 1.01	84.35 \pm 0.98
	Classifier (MIAFDR)	78.19 \pm 0.79	84.46 \pm 0.93
Tiny-ImageNet	Classifier	69.67 \pm 0.85	76.99 \pm 1.63
	Classifier (MIAFDR)	71.18 \pm 1.53	77.06 \pm 1.52

Table 1. Attack performance of classifier-based MIAFDR.

Further, we examine the performance of our proposed MIAFDR with likelihood ratio-based MIAs. Specifically, we evaluate the performance of the original LiRA [12] framework, which incorporates 64 shadow models, in comparison to its modified iteration, which integrates our MIAFDR approach. Figure 5a depicts the ROC curves of our proposed attacks and LiRA on CIFAR-10 using log scales. As we can see, our MIAFDR with LiRA approach yields superior log-scale ROC curves, exhibiting a higher True Positive Rate (TPR) at a lower False Positive Rate (FPR). Additionally, our MIAFDR with LiRA approach effectively manages FDR, as evidenced in Figure 5b. For instance, with $\pi_0 = 0.5$ and $\alpha = 0.15$, we achieve an empirical FDR of 0.145, which closely aligns with the analytical guarantee. Hence, our proposed attacks can be effectively integrated with likelihood ratio-based membership inference attacks to achieve both controlled FDR and good attack precision.

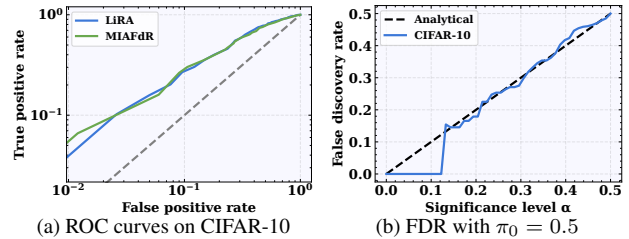


Figure 5. Attack performance of LiRA-based MIAFDR.

At last, we conduct experiments in the black-box setting, where the attacker lacks any prior knowledge about the private training dataset or the target pre-trained model, including its architecture and parameters. Here, we train the surrogate model in MIAFDR using an architecture that differs from the target model's. Figure 6a shows the attack

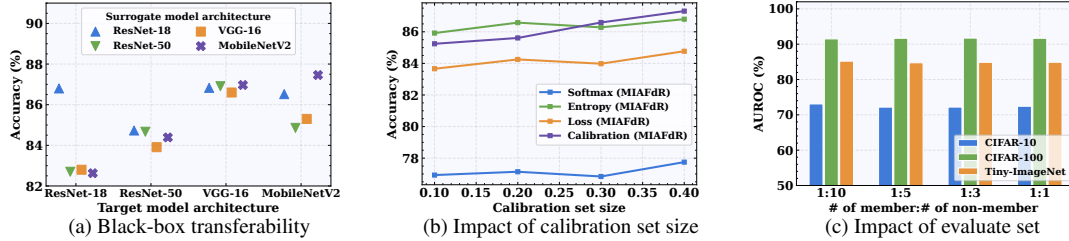


Figure 6. Black-box setting and ablation study of MIAFdr on CIFAR-10.

accuracy for various model architectures on Entropy-based MIAFdr. The reported experimental results in this figure demonstrate that our MIAFdr maintains robust attack performance across various model architectures, underscoring the effectiveness of our method in the black-box setting.

5.2. Ablation Study

We first perform an ablation study to explore the effectiveness of MIAFdr over calibration set size and the ratio of member to non-member data. We first investigate the impact of calibration set size on metric-based MIAFdr. As shown in Figure 6b, with a larger calibration set size, the attack accuracy tends to increase. This is because the inclusion of additional calibration data enhances the reliability of the non-member relative probability estimation, leading to more stable predictions in our MIAFdr. Next, we examine the impact of the evaluation set on MIAFdr. Figure 6c presents the AUROC score of MIAFdr across various ratios of member to non-member data. Remarkably, our method consistently exhibits a robust AUROC score irrespective of the ratio of member to non-member data in the evaluation set, thereby underscoring its consistent effectiveness.

derived experimental results in Figure 8a and Figure 8b. Specifically, Figure 8a indicates the effectiveness of our method in controlling the expected proportion of samples incorrectly reported to be memorized when they have actually been forgotten. Figure 8b presents the evaluation results regarding the effectiveness of lifelong learning in terms of memorizing data from previous tasks, as compared to evaluation based on accuracy. All of these experimental results verify the desired performance in traditional data memorization-based ML tasks with our proposed MIAFdr.

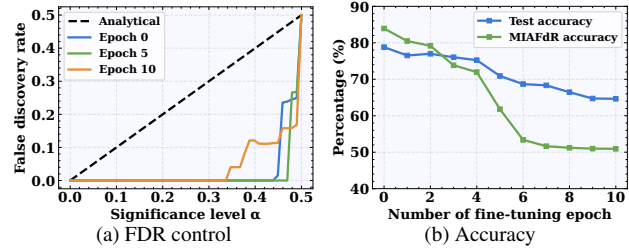


Figure 8. Lifelong learning with MIAFdr.

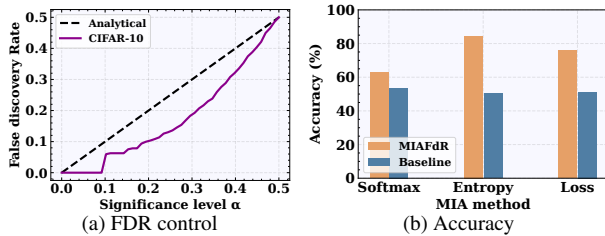


Figure 7. Machine unlearning with MIAFdr.

Additionally, we conduct ablation experiments to show the impact of our MIAFdr in enhancing traditional data memorization-based ML tasks. First, in Figure 7a, we report the obtained expected proportion results of instances erroneously identified as not unlearned in the context of machine unlearning. The results demonstrate that this proportion is effectively controlled across different significance levels. Here, we utilize a widely adopted popular unlearning method, i.e., SISA [11]. Figure 7b highlights that our method significantly outperforms the baselines in accuracy.

At last, we also evaluate our method in the lifelong learning task and adopt the fine tuning-based lifelong learning method [3]. For this lifelong learning setting, we report the

6. Conclusion

In this paper, we design a novel membership inference attack method, which can provide the false discovery rate guarantees. Notably, our proposed MIAFdr can work as a wrapper that can be seamlessly integrated with existing MIA methods in a post-hoc manner. Specifically, in our method, given the typically unknown true distributions of member and non-member data, we first design a novel conformity score function, which can reflect the conformity degree of test data to the non-member data. Then, based on the obtained point-wise conformity scores, we develop a non-member relative probability estimation strategy to assess the likelihood of not making discoveries. Following this, we introduce a novel adjustment method that modifies the initially estimated non-member relative probabilities to ensure the false discovery rate control, effectively addressing the challenges posed by interdependent non-member relative probabilities. We conduct the theoretical analysis for our method. Extensive experiments are conducted to verify the desired performance of our method. In particular, we also empirically show that our method can help data memorization-based ML tasks, including the unlearning verification task and the lifelong learning task.

Acknowledgements

This work is supported in part by the US National Science Foundation under grant CNS-2350332 and IIS-2442750. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 3
- [2] Chirag Agarwal, Daniel D’souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2022. 2
- [3] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. 8
- [4] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 5
- [5] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of statistics*, pages 2055–2085, 2015. 2
- [6] Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivald conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022. 2
- [7] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. 2, 5, 6
- [8] Yoav Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):405–416, 2010. 2
- [9] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. 2
- [10] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 6
- [11] Lucas Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021. 1, 8
- [12] 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. 1, 2, 3, 6, 7
- [13] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [14] Aobo Chen, Yangyi Li, Wei Qian, Kathryn Morse, Chenglin Miao, and Mengdi Huai. Modeling and understanding uncertainty in medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 557–567. Springer, 2024. 2
- [15] Aobo Chen, Yangyi Li, Chenxu Zhao, and Mengdi Huai. A survey of security and privacy issues of machine unlearning, 2025. 1
- [16] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 896–911, 2021. 1
- [17] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7766–7775, 2023. 1
- [18] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024. 1
- [19] Xiaoxiao Chi, Xuyun Zhang, Yan Wang, Lianyong Qi, Amin Beheshti, Xiaolong Xu, Kim-Kwang Raymond Choo, Shuo Wang, and Hongsheng Hu. Shadow-free membership inference attacks: Recommender systems are more vulnerable than you thought. *arXiv preprint arXiv:2405.07018*, 2024. 2
- [20] Gilad Cohen and Raja Giryes. Membership inference attack using self influence functions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4892–4901, 2024. 1, 2
- [21] Bruno De Finetti. *Theory of probability: A critical introductory treatment*. John Wiley & Sons, 2017. 5
- [22] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1
- [23] Ganesh Del Grosso, Hamid Jalalzai, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. Leveraging adversarial examples to quantify membership information leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10399–10409, 2022. 1, 2

- [24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [25] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 1
- [26] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020. 2
- [27] William Fithian and Lihua Lei. Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091–3118, 2022. 2
- [28] Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021. 2
- [29] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019. 1
- [30] Umang Gupta, Dimitris Stripelis, Pradeep K Lam, Paul Thompson, Jose Luis Ambite, and Greg Ver Steeg. Membership inference attacks on deep regression models for neuroimaging. In *Medical Imaging with Deep Learning*, pages 228–251. PMLR, 2021. 2
- [31] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European conference on computer vision*, pages 466–483. Springer, 2020. 2
- [32] 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)*, 2016. 1
- [33] 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*, pages 770–778, 2016. 6
- [34] Xinlei He, Hongbin Liu, Neil Zhenqiang Gong, and Yang Zhang. Semi-leak: Membership inference attacks against semi-supervised learning. In *European Conference on Computer Vision*, pages 365–381. Springer, 2022. 1, 2, 3
- [35] Yang He, Shadi Rahimian, Bernt Schiele, and Mario Fritz. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 519–535. Springer, 2020. 2
- [36] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008. 2
- [37] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022. 1, 2
- [38] Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3104–3122, 2021. 3
- [39] Adel Javanmard and Hamid Javadi. False discovery rate control via debiased lasso. 2019. 2
- [40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1
- [41] Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881, 2023. 1, 2
- [42] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders are scalable learners of cellular morphology. *arXiv preprint arXiv:2309.16064*, 2023. 1
- [43] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL <http://www.cs.toronto.edu/kriz/cifar.html>, 2009. 6
- [44] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023. 2
- [45] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 641–649, 2020. 6
- [46] Yangyi Li, Aobo Chen, Wei Qian, Chenxu Zhao, Divya Liddler, and Mengdi Huai. Data poisoning attacks against conformal prediction. In *International Conference on Machine Learning*, pages 27563–27574. PMLR, 2024. 2
- [47] Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for powerful out-of-distribution testing with labeled outliers. *arXiv preprint arXiv:2208.11111*, 2022. 2
- [48] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 521–534. IEEE, 2020. 2
- [49] Rong Ma, T Tony Cai, and Hongzhe Li. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, 116(534):984–998, 2021. 2

- [50] Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Machine learning meets false discovery rate. *arXiv preprint arXiv:2208.06685*, 2022. 2
- [51] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 77–83. IEEE, 2023. 2
- [52] Jun Niu, Peng Liu, Xiaoyan Zhu, Kuo Shen, Yuecong Wang, Haotian Chi, Yulong Shen, Xiaohong Jiang, Jianfeng Ma, and Yuqing Zhang. A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, 2024. 3
- [53] Wei Qian, Chenxu Zhao, Wei Le, Meiyi Ma, and Mengdi Huai. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1932–1942, 2023. 1
- [54] Wei Qian, Chenxu Zhao, Yangyi Li, Fenglong Ma, Chao Zhang, and Mengdi Huai. Towards modeling uncertainties of self-explaining neural networks via conformal prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14651–14659, 2024. 2
- [55] Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021. 2
- [56] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018. 6
- [57] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6
- [58] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR, 2021. 6
- [59] Avital Shafra, Shmuel Peleg, and Yedid Hoshen. Membership inference attacks are easier on difficult problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14820–14829, 2021. 2
- [60] Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9549–9557, 2021. 3
- [61] 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. 1, 3, 6, 7
- [62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6
- [63] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632, 2021. 6
- [64] Elham Tabassi, Kevin J Burns, Michael Hadjimichael, Andres D Molina-Markham, and Julian T Sexton. A taxonomy and terminology of adversarial machine learning. *NIST IR*, 2019:1–29, 2019. 1
- [65] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022. 1
- [66] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019. 2
- [67] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005. 2
- [68] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 1
- [69] Jiaqi Wang, Chenxu Zhao, Lingjuan Lyu, Quanzeng You, Mengdi Huai, and Fenglong Ma. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. *arXiv preprint arXiv:2407.03247*, 2024. 2
- [70] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. *Advances in neural information processing systems*, 32, 2019. 6
- [71] Zhen Wang, Liu Liu, Yiqun Duan, Yajing Kong, and Dacheng Tao. Continual learning with lifelong vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 171–181, 2022. 1
- [72] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021. 1
- [73] Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440*, 2021. 6
- [74] Tianxiang Xu, Chang Liu, Kun Zhang, and Jianlin Zhang. Membership inference attacks against medical databases. In *International Conference on Neural Information Processing*, pages 15–25. Springer, 2023. 2
- [75] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022. 2
- [76] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer se-*

- curity foundations symposium (CSF)*, pages 268–282. IEEE, 2018. [2](#), [6](#)
- [77] Soroush H Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. *arXiv preprint arXiv:2407.09165*, 2024. [2](#)
- [78] Sajjad Zarifzadeh, Philippe Cheng-Jie Marc Liu, and Reza Shokri. Low-cost high-power membership inference by boosting relativity. *arXiv preprint arXiv:2312.03262*, 2023. [1](#)
- [79] Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. Label-only membership inference attacks and defenses in semantic segmentation models. *IEEE Transactions on Dependable and Secure Computing*, 20(2):1435–1449, 2022. [2](#)
- [80] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. Membership inference attacks against recommender systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 864–879, 2021. [2](#)
- [81] Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, and Yang Zhang. Generated distributions are all you need for membership inference attacks against generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4839–4849, 2024. [2](#)
- [82] Rongting Zhang, Martin Bertran, and Aaron Roth. Order of magnitude speedups for llm membership inference. *arXiv preprint arXiv:2409.14513*, 2024. [3](#)
- [83] Chenxu Zhao, Wei Qian, Rex Ying, and Mengdi Huai. Static and sequential malicious attacks in the context of selective forgetting. *Advances in Neural Information Processing Systems*, 36:74966–74979, 2023. [1](#)
- [84] Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5713–5723, 2023. [1](#)
- [85] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. [1](#)