Enhanced Label-Only Membership Inference Attacks with Fewer Queries

Hao Li*
Institute of Software, Chinese Academy of Sciences

Zheng Li*
Shandong University

Siyuan Wu
Institute of Software, Chinese Academy of Sciences

Yutong Ye
Institute of Software, Chinese Academy of Sciences

Min Zhang†
Institute of Software, Chinese Academy of Sciences

Dengguo Feng
Institute of Software, Chinese Academy of Sciences

Yang Zhang
CISPA Helmholtz Center for Information Security

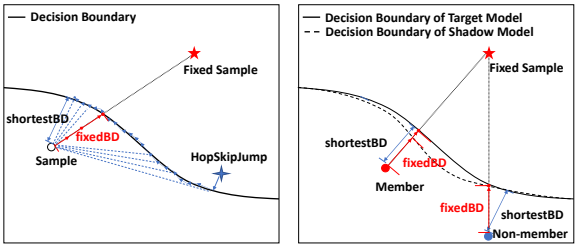
Abstract

Machine Learning (ML) models are vulnerable to membership inference attacks (MIAs), where an adversary aims to determine whether a specific sample was part of the model's training data. Traditional MIAs exploit differences in the model's output posteriors, but in more challenging scenarios (label-only scenarios) where only predicted labels are available, existing works directly utilize the shortest distance of samples reaching decision boundaries as membership signals, denoted as the shortestBD. However, they face two key challenges: low distinguishability between members and non-members due to sample diversity, and high query requirements stemming from direction diversity.

To overcome these limitations, we propose a novel label-only attack called `DHAttack`, designed for Higher performance and Higher stealth, focusing on the boundary distance of individual samples to mitigate the effects of sample diversity, and measuring this distance toward a fixed point to minimize query overhead. Empirical results demonstrate that `DHAttack` consistently outperforms other advanced attack methods. Notably, in some cases, `DHAttack` achieves more than an order of magnitude improvement over all baselines in terms of TPR @ 0.1% FPR with just 5 to 30 queries. Furthermore, we explore the reasons for `DHAttack`'s success, and then analyze other crucial factors in the attack performance. Finally, we evaluate several defense mechanisms against `DHAttack` and demonstrate its superiority over all baseline attacks.[1]

1 Introduction

Over the past decade, machine learning (ML) has significantly influenced various human activities, including item recommendation for online shopping [14, 19, 36], route planning for vehicles [4, 41], and handwriting recognition [5, 30, 31] for delivery operations. ML models assist in decision-making by offering predicted outcomes, often with associated confidence levels, such as predicted probabilities. These models have profoundly affected both work and daily life.



(a) Number of Queries (b) Distinguishability

Figure 1: An illustration of the limitations of shortestBD and the intuitions behind our method.

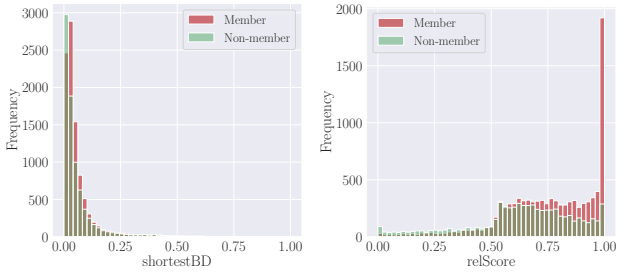
Meanwhile, ML models also face privacy attacks. Membership inference attacks (MIAs) are among the most severe in the ML domain, aiming to determine whether a sample was used in a model's training. Most existing studies [10, 18, 34, 37, 39, 47] exploit differences in the target model's output posteriors for members and non-members to conduct their attacks. In more realistic and challenging scenarios where models return only predicted labels rather than posterior probabilities, several studies [11, 26, 43, 44, 48] have demonstrated that membership leakage still occurs. These studies highlight a key insight: members of the training set tend to be positioned farther from the decision boundary compared to non-members. We denote the distance to the decision boundary as the *boundary distance* (BD). Building on this insight, existing label-only attacks [11, 26] often measure the minimal perturbations—calculated through adversarial examples—required to change a sample's predicted label. If the minimal perturbation exceeds a certain threshold, the sample is classified as a member; otherwise, a non-member. In this paper, we refer to these minimal perturbations as the *shortest boundary distance* (shortestBD).

*The first two authors made equal contributions.
†Corresponding author.
[1]Code is available at https://github.com/AIPAG/DHAttack or https://doi.org/10.5281/zenodo.14728863.

However, employing the shortestBD for attacks has two main limitations: *high query requirement* and *low distinguishability*. First, due to *direction diversity*, identifying a sample's shortestBD requires querying from various directions, which demands a large number of queries. For example, existing label-only attacks [11, 26, 44, 48] that employ advanced adversarial example techniques, such as HopSkipJump [9] or QEBA [24], typically need over 1,000 queries to identify the shortestBD, as shown in Figure 1a. This approach is costly for adversaries and can be easily detected by model owners. More importantly, the assumption that members consistently have larger shortestBD values than non-members is somewhat idealized. Due to *sample diversity*, some non-members inevitably exhibit shortestBD values similar to those of members, as shown in Figure 1b. This overlap in the shortestBD distributions of members and non-members, illustrated in Figure 2a, reduces attack performance and increases false positive rates.

To address these problems, we propose a novel membership signal that also relies on boundary distance but from two new perspectives. First, instead of identifying a sample's shortestBD among various directions, we measure the distance at which a sample crosses the decision boundary when moved toward a fixed data sample, denoted as fixedBD, as shown in Figure 1a. The fixed sample acts as a consistent reference point, streamlining computation by eliminating the need for costly shortestBD searches and ensuring uniformity in boundary distance measurements. Second, instead of comparing the boundary distances between different samples, we shift our focus to each individual sample: a sample will exhibit a larger boundary distance if it was in the training set compared to if it itself was not. This can reduce the negative effect of the overlap between member and non-member boundary distances, thereby potentially achieving better attack performance.

Building on the above, we propose a new label-only membership inference attack called `DHAttack`, designed for Higher performance and Higher stealth. The adversary first trains several local models (called shadow models) using a dataset drawn from the same distribution as the target model's training set, but without including the target samples being inferred. For the target sample, the adversary measures its boundary distances toward the fixed data sample (e.g., an image with all RGB values set to 255) on each shadow model, thereby generating a set of fixedBDs, which is then modeled as a Gaussian distribution. Since the target sample is definitely not included in the shadow models' training, this distribution actually represents its non-member state. As shown in Figure 1b, although member and non-member samples may exhibit similar shortestBD and fixedBD values, a member's fixedBD on the target model is greater than on the shadow model. In contrast, non-members display consistent fixedBD values across both models. Therefore, the adversary measures the target sample's fixedBD on the target victim model and computes the CDF (Cumulative Distribution Function) value



(a) ShortestBD Distribution (b) RelScore Distribution

Figure 2: Distributions of shortestBD (normalized) and relScore for members/non-members. Frequency means the number of samples. The model is ResNet-56 with CIFAR10.

of this fixedBD over its non-member state distribution. A larger CDF value indicates that the fixedBD of the target sample on the target model is greater than most of the fixedBDs drawn from this distribution. This suggests a higher likelihood of the sample being a member. Conversely, a smaller CDF value implies a higher likelihood of the sample being a non-member. Thus, the CDF value can serve as the membership signal, denoted as the relative membership score (relScore) in this work. This new signal effectively enhances the distinction between members and non-members, as illustrated in Figure 2b.

We conduct extensive evaluations on five benchmark datasets and compare `DHAttack` with existing label-only methods. The results show that `DHAttack` consistently outperforms the baselines in nearly all cases. For example, when using MobileNetV2 trained on CIFAR10, `DHAttack` achieves more than an order of magnitude improvement over all baselines in terms of TPR @ 0.1% FPR with just 5 to 30 queries. Additionally, we explore why `DHAttack` achieves higher performance with fewer queries and present ablation studies to evaluate the impact of various factors on its effectiveness. Finally, we evaluate `DHAttack` and other baselines against several representative defenses. The results demonstrate that `DHAttack` delivers the best performance in almost all scenarios, particularly in TPR @ 0.1% FPR. Overall, our contributions can be summarized as follows:

- We introduce a new membership signal for label-only scenarios with two key insights: (1) focusing on the state of the samples themselves, reducing the negative impact of sample diversity, and (2) measuring the boundary distance to a fixed point, reducing the large query requirements for shortestBD.

- We propose a new label-only attack, `DHAttack`, which leverages the newly introduced membership signal, called relScore. Empirical results show that `DHAttack` consistently outperforms baseline methods, achieving over a tenfold improvement in TPR @ 0.1% FPR with

fewer than 50 queries in certain cases.

- We explore the reasons for DHAttack's success, conduct comprehensive ablation studies of key factors, and evaluate its performance against representative defenses.

2 Preliminaries

2.1 Label-Only MIAs

Membership inference attacks aim to determine whether a target data sample is used in a target model's training. Here, we present a formal definition of MIAs in the label-only scenario.

Definition 1. Given a data sample x with its class label $y \in C$, a trained ML model f_T and some auxiliary information of an adversary, denoted as I. Then, the membership inference attack $\mathcal{A}$ can be defined as follows:

$$f_T : x \to \hat{y} \in C.$$

$$\mathcal{A} : x, y, f_T, I \to \{0, 1\}.$$

Here, 0 means x is not a member of f_T's training set, and 1 otherwise.

2.2 Threat Model

In the label-only setting, the adversary can only access the target model's predicted labels without any posteriors. Furthermore, we follow two assumptions about the adversary's training knowledge in the prior MIAs [11,27,34,37,39,44–46,48]. As [11] states, this training knowledge "could be publicly available or inferable from a model extraction attack."

Same Data Distribution. The adversary holds an auxiliary dataset drawn from the same distribution as the target model's training set.

Same Model Architecture. The adversary has the knowledge of the architecture and hyperparameters of the target model.

Since these two assumptions significantly influence our approach, we further relax them in Section 5.4. Additionally, a successful attack must meet the following requirement, which most previous label-only attacks [11,26,44,48] fail to satisfy, rendering them largely inapplicable.

Limited Queries. The number of queries to the target model should be minimized. This limitation is necessary because queries are typically charged per request, and a high query volume in a short timeframe may trigger security alerts.

2.3 Boundary Distance and Membership Signal

Boundary distance serves as the primary signal for prior label-only MIAs [11,26,44,48]. This section will formally define it to clarify the content of the subsequent sections. Furthermore, we provide a formal definition of our membership signal, relScore.

Definition 2 - ShortestBD. ShortestBD is the minimum distance from a sample x to the decision boundary of a model f_T. Using an adversarial perturbation algorithm A_G and a query budget K, this value can be computed as follows:

$$A_G : x, K \to \hat{x}, \text{ subject to } f_T(x) \neq f_T(\hat{x});$$

$$\text{shortestBD} : x, \hat{x} \to ||\hat{x} - x||_2.$$

Here, $\hat{x}$ refers to the smallest perturbation sample found by A_G under the constraint of K queries. Typically, it requires thousands of queries to find an approximately minimal perturbation, which has been validated in prior MIAs [11,26,44,48].

Definition 3 - FixedBD. FixedBD is the distance that the sample x travels toward the fixed point x_{fixed} until it reaches the model decision boundary. Given x, its ground truth label y, a model f_T, and a query budget K, this value can be computed as follows:

$$x_{\text{diff}} = x_{\text{fixed}} - x$$

$$\text{fixedBD} : x, y, x_{\text{diff}}, K \to \arg\min_{i \in [0,K]} f_T(x + \frac{i}{K} \cdot x_{\text{diff}}) \neq y.$$

Here, we typically select outliers (samples with few neighbors) from the training data distribution of f_T as x_{fixed} to ensure that most samples cross the decision boundary. In our DHAttack, the membership signal is derived by comparing the fixedBD of a sample in the target model with that in shadow models. Consequently, we only require a consistent x_{fixed}, and the need for shortest distance optimization is eliminated, thereby reducing the number of queries.

Definition 4 - RelScore. In this paper, we adopt relScore, derived from fixedBD, as the membership signal. Given x's fixedBD value d on the target model and $\{d_1, ..., d_n\}$ from n shadow models excluding x from training, the relScore is computed as follows:

$$\{d_1, ..., d_n\} \to G$$

$$\text{relScore} : d, G \to \text{CDF}(d, G).$$

Here, G denotes the Gaussian distribution constructed from $\{d_1, ..., d_n\}$, and CDF refers to its Cumulative Distribution Function.

3 Attack Methodology

3.1 Design Intuition

As aforementioned, existing label-only MIAs confront the problems of limited attack performance and large queries to the target model. This motivates us to find the reasons behind these problems.

Intuition 1. First, we explore whether the distinguishability of boundary distances related to membership can be improved.

Table 1: Query numbers for measuring boundary distance using different methods (on VGG-16 with CIFAR10).

Sample	Status	fixedBD		shortestBD	
		# Queries	Norm. Dist.	# Queries	Norm. Dist.
A	Member	10	0.800	100	0.621
		20	0.750	500	0.604
		30	0.733	1000	0.462
		50	0.720	2000	0.481
B	Non-member	10	0.300	100	0.652
		20	0.300	500	0.597
		30	0.267	1000	0.115
		50	0.260	2000	0.103
C	Member	10	0.700	100	0.715
		20	0.700	500	0.426
		30	0.667	1000	0.490
		50	0.660	2000	0.387
D	Non-member	10	0.200	100	0.729
		20	0.200	500	0.570
		30	0.167	1000	0.109
		50	0.160	2000	0.115

Recent studies [7,42,45] have shown that membership leakage risk varies across samples, partly because some members and non-members exhibit similar losses. Similarly, in the label-only scenario, we hypothesize that some members and non-members also have similar boundary distances, as illustrated in Figure 2a and later evaluations.

To address this, our first attack strategy aims to mitigate the negative impact of sample diversity on boundary distance. Specifically, we compare each target sample's boundary distance in two states—whether it participated in model training or not—under the premise that a sample will have a larger boundary distance if it was in the training set compared to if it was not. Inspired by LiRA [7], the simplest method is to train in-models and out-models locally to estimate a sample's boundary distance in both member and non-member states. However, as highlighted in [7, 17, 23, 27], training in-models necessitates retraining for each new target sample (or batch), making it impractical in real-world settings. Furthermore, He et al. [17] note that LiRA's implementation [7] assumes the adversary has simultaneous access to all target samples, constructing in-models and out-models for the entire dataset at once. This reduces the MIA problem to ranking samples within a fixed dataset, deviating from standard MIA scenarios.

To avoid building two states for a target sample, we only approximate its non-member state using multiple local shadow models trained on data excluding the target sample, where the obtained boundary distances are modeled as Gaussian distributions. The target sample's actual boundary distance is then measured on the target model, and membership is inferred by computing the cumulative distribution function (CDF) of this distance within the Gaussian distribution. This CDF value is called the relative membership score (relScore), with higher values indicating that the actual boundary distance exceeds most of the boundary distances in the non-membership state, thus suggesting a greater likelihood of being a member.

Intuition 2. Our second focus is on reducing the number of queries required to the target model. In label-only scenarios, previous works [11,26,44,48] often use the shortest boundary distance (shortestBD)—the distance between the target sample and its adversarial example—as the primary membership signal. However, accurately determining this distance across various directions typically requires a large number of queries (often exceeding 1,000).

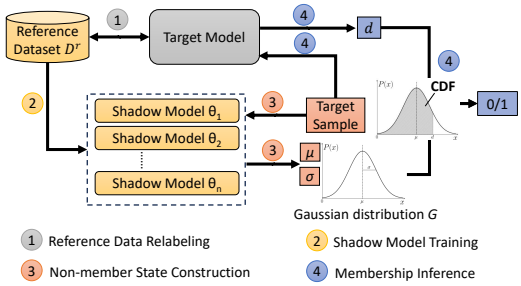


Figure 3: Overview of DHAttack.

To minimize query costs, our second attack strategy measures the distance at which a sample crosses the decision boundary when moved toward a fixed data point, which is called the fixed boundary distance (fixedBD). This approach eliminates the need for extensive queries required to compute the shortestBD across multiple directions. In our implementation, we select an out-of-distribution data point relative to the target samples. If the fixed data point still belongs to the same distribution as the target samples, there will likely be a significant number of samples that do not cross the decision boundary, as they may be neighboring samples of the fixed point. Therefore, for simplicity, we consider an out-of-distribution data point: an image with all RGB values set to the maximum value of 255. This ensures that all target samples can cross the decision boundary toward it.

Table 1 compares the query costs of several examples between measuring fixedBD and shortestBD. We find that accurately measuring the shortestBD requires at least 1,000 queries, as using fewer queries (e.g., 100 or 500) yields unreliable results. For example, with 500 queries, the shortestBD for Member C (0.426) is smaller than that for Non-member D (0.570), contradicting the results obtained with 1,000 and 2,000 queries. In contrast, fixedBD requires less than 30 queries for accurate measurement.

3.2 Attack Pipeline

Based on the above, we propose a new label-only membership inference attack, namely DHAttack, which aims to achieve Higher performance and Higher stealth. The pipeline of DHAttack is depicted in Figure 3, involving four phases: reference data relabeling, shadow model training, non-member state construction, and membership inference. See Algorithm 1 in Appendix B.

Reference Data Relabeling. As mentioned in Section 2.2,

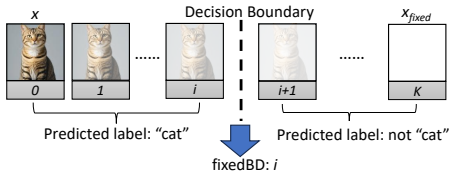


Figure 4: The calculation of fixedBD. The intermediate samples, indexed 1 to K, are equidistant points from the target sample x to the fixed sample x_{fixed}.

the adversary possesses an auxiliary dataset, referred to as the reference dataset D^r, which shares the same distribution as the target model's training set. The adversary first queries the target model using samples from D^r and relabels this dataset based on the target model's predictions. This relabeled reference dataset captures the predictive behavior of the target model, enabling the local model trained on it to approximate the target model's decision boundary.

Note that the relabeling query for the reference dataset is performed only once during the attack process. The number of queries required for this relabeling is acceptable for the adversary. For instance, in our experiments, the default reference dataset size is 20,000, requiring the same number of queries as those needed to infer just 20 target samples using baseline methods like SBA [11] and UBA [26]. Additionally, these relabeling queries can occur during normal usage of the target model, with the adversary acting as a regular user. As a result, this activity is unlikely to be detected by the model owner.

Even though the relabeling queries are feasible and realistic, we further investigate the possibility of removing the relabeling operation entirely, as discussed in Section 5.4.

Shadow Model Training. The adversary can then leverage the relabeled dataset D^r to train n local models, referred to as shadow models $(\theta_1, \theta_2, \ldots, \theta_n)$. Each shadow model represents a decision boundary learned without target samples, but it is similar to the decision boundary of the target model. Note that our second assumption is that these shadow models share the same architecture and hyperparameters as the target model. Therefore, these shadow models can be used to approximate the non-member status of a target sample on the target model. However, we also acknowledge that the second assumption is overly strict, limiting its applicability in real-world settings. We further discuss its relaxation in Section 5.4.

Non-member State Construction. We now approximate the non-member state of the target sample x. It is important to note that the target sample is not included in the training of the shadow models; therefore, its behavior on these shadow models can be regarded as the sample's non-member state.

The adversary begins by selecting a fixed data sample that lies outside the distribution of the target sample, such as an image with all RGB values set to 255 (denoted as "RGB-255"). Then, we propose a simple yet effective method to determine the boundary distance of a target sample using equal-sized steps. Specifically, the adversary computes the difference between the target sample x and the fixed sample x_{fixed} as $x_{fixed} - x$. This difference is divided into K equidistant parts and incrementally added to the target sample x, generating intermediate samples with index $i \in \{1, ..., K\}$. The fixedBD is determined at the index i where the intermediate sample retains the predicted label, but the addition at index $i+1$ causes a label change. Figure 4 illustrates this process. See Appendix B Algorithm 2 for details.

Here, we emphasize two critical aspects. First, the complexity of decision boundaries in high-dimensional spaces can cause classification inconsistencies among the K queries between the target and fixed samples, leading to inaccurate fixedBD estimates. While increasing K enhances the accuracy, it also magnifies the effects of these inconsistencies, ultimately lowering attack performance (see Figure 5). Based on our experiments in Section 5.1, we recommend selecting K from 10 to 50, which we find sufficient for our attack. Besides, although binary search is more efficient, it is highly sensitive to boundary complexity and is therefore not used in this process. Second, choosing an appropriate fixed sample, such as an outlier (e.g., "RGB-255"), is crucial. In high-dimensional spaces, non-neighboring samples of an outlier typically cross the decision boundary when moving toward it. Only a few neighboring samples may fail to cross the boundary, in which case their fixedBDs are directly set to K. See Section 5.2 for more details.

Overall, for a target sample, the adversary repeats the process across all n local shadow models, obtaining n fixedBD values. These values are modeled as a Gaussian distribution G. Since the target sample is excluded from the training datasets of the shadow models, this Gaussian distribution represents the target sample's non-member state.

Membership Inference. Similarly, the adversary feeds the target sample x and its K intermediate samples into the target model to obtain the actual fixedBD, denoted as d. Using this fixedBD d and the non-member state distribution G, the adversary calculates the Cumulative Distribution Function (CDF) value of d, which is used as a membership score, referred to as the relative membership score (relScore). A larger relScore indicates that the target sample's fixedBD on the target model exceeds most fixedBDs drawn from G, suggesting a higher likelihood of the sample being a member. Conversely, a smaller relScore implies a greater likelihood of the sample being a non-member.

In implementation, we select a threshold τ for membership inference. If the relScore exceeds τ, it is classified as a member; otherwise, it is a non-member. This can be formulated as the indicator function $\mathbb{I}[relScore > \tau]$. Following [7, 29, 45], the adversary can explore a range of threshold values to obtain the trade-off between TPR and FPR. In our method, a shadow model, whose members and non-members are known to the adversary, can be used to identify τ to reach the specified FPR.

Table 2: Training/testing accuracy of all target models in our experiments.

Target model	CIFAR10	CIFAR100	CINIC10
VGG-16	1.000/0.756	1.000/0.296	1.000/0.569
ResNet-56	0.987/0.662	0.998/0.243	0.972/0.472
MobileNetV2	0.986/0.667	0.998/0.218	0.972/0.463
Target model	Purchase	News	
MLPs	1.000/0.716	0.976/0.663	

We also propose a simple threshold selection method that does not require a shadow model in Section 5.3.

4 Experimental Setup

4.1 Datasets

We evaluate `DHAttack` on three image datasets and two non-image datasets, namely CIFAR10 [22], CIFAR100 [22], CINIC10 [12], Purchase [1], and News [2]. These datasets are commonly used in existing label-only membership inference attacks [11, 26, 43, 44, 48]. See details in Appendix A.

Given the varying requirements of different baselines, we follow [27] to randomly split each dataset into five disjoint parts: target training/testing dataset (D^t_{train} / D^t_{test}), shadow training/testing dataset (D^s_{train} / D^s_{test}), and reference dataset D^r. The former two sets are held by the target model's owner, which are treated as members and non-members of the target model. The latter three sets are held by the adversary. Among them D^s_{train} and D^s_{test} are used by NRA [11], SBA [11] and TrajectoryMIA [27] to train and test a shadow model, which are treated as members and non-members of the shadow model. While D^r is used by TrajectoryMIA [27], YOQO [43] and our `DHAttack` to train several local models, as described in these works. The number of samples in each part is described in Appendix A Table 11.

4.2 Models

We use three popular model architectures for image datasets: VGG-16 [38], ResNet-56 [16], and MobileNetV2 [35]. For non-image datasets, we follow prior work [27] to adopt a 2-layer MLP as the target model. All models are trained using the SGD algorithm, with training lasting between 80 and 150 epochs and learning rates ranging from 0.01 to 0.1, depending on the model's complexity. The performance of the target models is shown in Table 2. These same architectures and hyperparameters are also used for the shadow models trained by the adversary.

4.3 `DHAttack` Hyperparameters

First, the number of shadow models is 256 by default. Note that the adversary only needs to train these shadow models once in the whole attack process, and thus the computational cost is tolerable. Additionally, the number of segments K is set to 30 by default. Note that we also study its impact on the attack performance by varying it.

4.4 Baselines

In this paper, we adopt the following label-only MIAs as baselines.

Noise Robustness Attack (NRA). NRA is a method proposed by Choquette-Choo et al. [11], which adds Gaussian noise to the target sample x to obtain K noisy samples, and calculates the target model's prediction accuracy of the K samples to approximate x's boundary distance. The K noisy samples mean that the adversary has to launch K queries to the target model for each target sample. Commonly, the accuracy of K noisy samples from members exceeds that of non-members.

Unsupervised Boundary-attack (UBA). UBA [26] is a label-only MIA using the adversarial attack algorithms HopSkipJump [9] to find an adversarial example e of the target sample x. Then, the distance between x and e is taken as x's boundary distance. Their experiments demonstrate that members usually have a lager boundary distance than non-members. However, it would take more than 1,000 queries on the target model to find x's adversarial example and launch a successful attack. Furthermore, they compute several random samples' boundary distances and take the top t percentile over these distances as the threshold to distinguish members and non-members.

Supervised Boundary-attack (SBA). SBA is a similar label-only attack to UBA, proposed by Choquette-Choo et al. [11]. The only difference is the threshold-choosing method. The adversary of SBA launches the attack against the shadow model and finds the best threshold to distinguish members and non-members of the shadow model. Finally, this threshold is used to attack the target model.

TrajectoryMIA for Label-only (TrajectoryMIA). Liu et al. [27] presents a state-of-the-art MIA for the common black-box scenario. In this paper, we use its label-only version as a baseline. First, the adversary mimics the training process of the target model by distillation and obtains one sample x's multiple losses from intermediate model versions of the distilled model. Second, the adversary adopts HopSkipJump [9] to get x's boundary distance. Then, the losses and boundary distance of x are constructed as a feature vector. Following the approach, the adversary can construct many feature vectors to train an MLP-based attack model. Finally, the well-trained attack model can be used to infer the membership of the target sample.

YOQO. YOQO [43] is an alternative attack method that does not use boundary distance, and focuses on reducing the number of queries to the target model. The adversary trains N in-models (trained with the target sample x), and N out-models (trained without x). Then the adversary crafts a query sam-

ple x', whose cross-entropy losses with x's ground-truth label c_x are small on the in-models and large on the out-models. Finally, x' is fed to the target model. If the predicted class label of x' is the same as c_x, x is inferred as a member, otherwise non-member. This attack only queries once to the target model but only achieves similar performance to SBA.

4.5 Metrics

We adopt the following two evaluation metrics.

AUC. AUC is an average-case metric widely used in earlier studies [10, 26, 34, 47]. It is the area under the receiver operating characteristic (ROC) curve, which is suitable for binary classification tasks, such as membership inference attacks.

TPR @ low FPR. TPR @ low FPR (True-Positive Rate at low False-Positive Rate) is a novel metric for evaluating MIAs, proposed by Carlini et al. [7], which reports the true-positive rate at a single low false-positive rate (e.g., 0.1% FPR). They argue that the high attack performance at high FPR is useless for the adversary, and a reliable inference attack targeting a small portion of the dataset is more valuable. Consequently, the metric is adopted in many recent MIA works [23, 27, 44, 45] to evaluate the reliability of an attack.

5 Experimental Results

5.1 Attack Performance

Figure 5 and Figure 6 present the TPR @ 0.1% FPR and the AUC, respectively. See more results in Appendix Figure 18 and Figure 19. `DHAttack` achieves optimal performance within 100 queries, so we limit our tests to a maximum of 200 queries. For UBA and SBA, we evaluate performance from 125 queries up to more than 20,000 queries, as these methods require a large number of queries to determine the shortestBD. Note that for UBA and SBA, we control the number of queries by limiting the maximum number of iterations in the HopSkipJump algorithm to find an adversarial example. With a maximum of 1 iteration, the number of queries is 125, which is why we do not test with fewer queries for these methods. For NRA, we use hundreds of queries to achieve optimal performance.

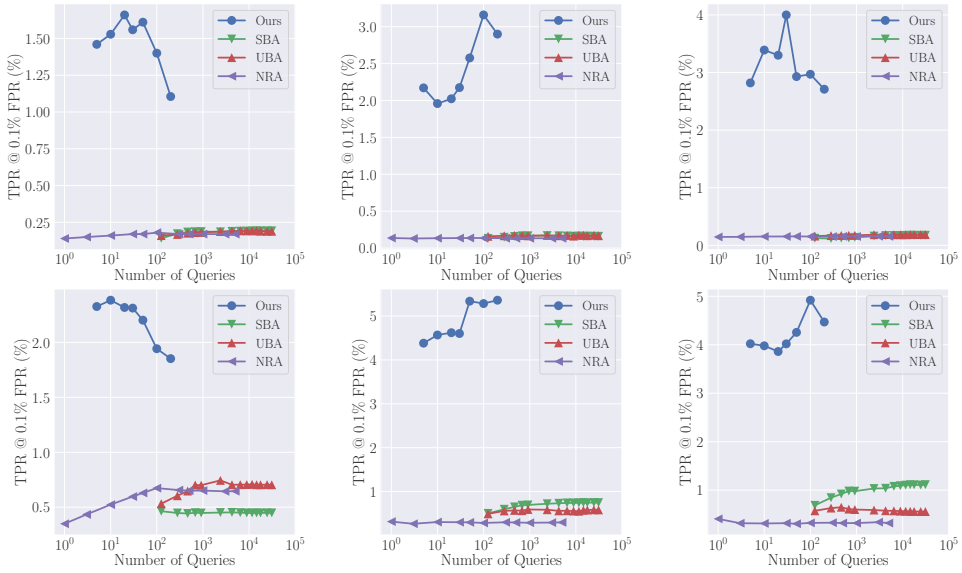
First, we focus on query costs, as the performance of NRA, UBA, and SBA heavily depends on the number of queries. For TrajectoryMIA, which requires training new attack models for different query counts, we fix the query count at 1,000 in our experiments. While YOQO needs only one query on the target model, it requires crafting a query sample x' across all in-models and out-models, making it computationally expensive. Therefore, TrajectoryMIA and YOQO are excluded from query cost comparisons. As shown in Figure 5 and Figure 6, `DHAttack` consistently outperforms NRA, UBA, and SBA, achieving effective results with very few queries, whereas the baseline methods require over 100 queries to perform effectively. Additionally, we report the best performance of these attacks, including the number of queries required to achieve optimal results, along with the performance of TrajectoryMIA and YOQO, as shown in Table 3. We can observe that, in some cases, `DHAttack` shows an order of magnitude improvement over other baselines. For example, for ResNet-56 trained on CIFAR-10, `DHAttack` achieves 2.58% TPR @ 0.1% FPR, while the baselines fall below 0.2%. Moreover, `DHAttack` achieves optimal performance within 50 queries, while UBA, SBA, and TrajectoryMIA require at least 1,000 queries yet deliver subpar performance. Although NRA requires several hundred queries and YOQO only one, both demand fewer queries than UBA, SBA, and TrajectoryMIA, but their performance consistently ranks among the lowest. Lastly, we note that `DHAttack` achieves an AUC of 0.719 on VGG-16, slightly lower than the best baseline TrajectoryMIA's 0.730. However, its TPR @ 0.1% FPR is nearly 5 times higher than that of TrajectoryMIA. More importantly, `DHAttack` requires only 30 queries to the target model, compared to the 1000 queries needed by TrajectoryMIA to achieve its best results. Therefore, when considering all metrics, including the number of queries, `DHAttack` significantly outperforms all baselines.

In summary, `DHAttack` achieves superior performance with significantly fewer queries. This advantage is attributed to the relScore used by `DHAttack`—a sample-dependent membership signal that indicates the likelihood that a sample's actual distance, measured in a fixed direction, exceeds the values drawn from the approximated non-member state distribution.

5.2 Analysis

Effect of Measurement Directions. As shown in Figure 7, the boundary distances for both member and non-member samples vary significantly across 100 random directions. Additionally, using the shortest boundary distance found in these directions for membership inference reveals only a modest difference between members and non-members. This finding supports the conclusion in [48] that using the shortest boundary distance as a membership signal may not yield effective attack performance. Consequently, existing label-only MIAs based on adversarial attack algorithms [11, 26, 27, 44] often expend many queries on the target model to identify a suboptimal membership signal, i.e., the shortest boundary distance.

To reduce unnecessary queries to the target model, we use a single fixed point to measure the boundary distance (i.e., fixedBD) instead of pursuing the shortest distance. As mentioned in Section 3.1, if the fixed point has many neighbors, these neighbors may not cross the boundary, regardless of their involvement in training. Therefore, we adopt an out-of-distribution fixed point to ensure that most target samples can cross the boundary and obtain a discriminative fixedBD. We further validate this intuition by attacking Mo-



(a) VGG-16 (b) ResNet-56 (c) MobileNetV2

Figure 5: TPR @ 0.1% FPR under different numbers of queries for attacks on three model architectures and two image datasets (from top to bottom: CIFAR10 and CIFAR100).

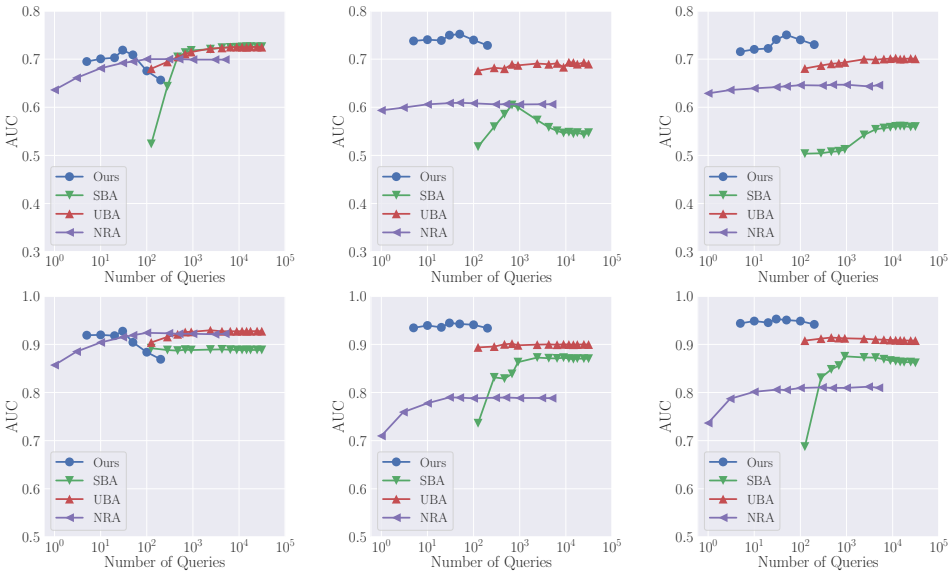
Table 3: The best attack performance of `DHAttack` and baselines against three target models trained on CIFAR10.

MIA method	TPR @ 0.1% FPR (%)			AUC		
	VGG-16	ResNet-56	MobileNetV2	VGG-16	ResNet-56	MobileNetV2
NRA	0.17(0.3k)	0.14(0.1k)	0.15(1k)	0.700(0.3k)	0.608(0.1k)	0.647(1k)
UBA	0.19(21k)	0.17(0.7k)	0.17(15k)	0.726(21k)	0.605(0.7k)	0.561(15k)
SBA	0.19(6.5k)	0.17(11k)	0.18(11k)	0.725(6.5k)	0.694(11k)	0.702(11k)
TrajectoryMIA	0.34(1k)	0.14(1k)	0.17(1k)	**0.730(1k)**	0.615(1k)	0.642(1k)
YOQO	0.18(1)	0.18(1)	0.17(1)	0.718(1)	0.717(1)	0.696(1)
DHAttack	**1.56(30)**	**2.58(50)**	**2.93(50)**	0.719(30)	**0.752(50)**	**0.750(50)**

bileNetV2 trained on CIFAR10. We randomly selected 100 samples from GTSRB [40], a widely used traffic sign dataset (see details in Appendix A), as fixed points for launching `DHAttack`, referred to as "Outside." Additionally, we sampled 100 points from target samples (i.e., CIFAR10) as fixed points for `DHAttack`, referred to as "Inside." Moreover, we also used an image with all RGB values set to 255 as a special outside fixed point (denoted as "RGB-255"). As shown in Figure 8, we observe that `DHAttack` (Outside) performs better than `DHAttack` (Inside), indicating that choosing fixed points outside the distribution of target samples is more effective from an attack perspective. Furthermore, RGB-255, which is located at the edge of the target sample distribution as shown in Figure 9, has fewer neighbors, and the attack performance using it surpasses that of most other `DHAttack` (Outside) points. Consequently, RGB-255 is used as the default fixed point in our experiments.

Effect of Samples' Diversity. As shown in Figure 10, we measure the boundary distances (fixedBD) of 10,000 members and 10,000 non-members using a fixed out-of-distribution data sample. The distance distributions for members and non-members show substantial overlap. This overlap is due to the diversity of samples; non-members with easily learned features also have large boundary distances, making it challenging for an attacker to differentiate them from members. This explains the high false positive rates of many existing methods that use raw distances for attacks. See more results in Appendix Figure 17.

To address the challenge posed by sample diversity, we construct a non-member state for the target sample by creating a Gaussian distribution of its fixedBDs, measured across several local shadow models trained without the target sample. We then use the CDF value of the target sample's actual fixedBD (obtained from the target model) over the Gaussian distribution as a membership signal, i.e., relScore. As shown in Figure 10, using relScore to distinguish between members and non-members improves discrimination in samples with large relScore values, resulting in a high TPR at a low FPR. Figure 11 further shows the Log-scale ROC curves for distinguishing members and non-members using fixedBD and relScore in the three scenarios from Figure 10 and Appendix Figure 17. By using non-member status as a refer-



(a) VGG-16 (b) ResNet-56 (c) MobileNetV2

Figure 6: AUC under different number of queries for attacks on three model architectures and two image datasets (from top to bottom: CIFAR10 and CIFAR100).

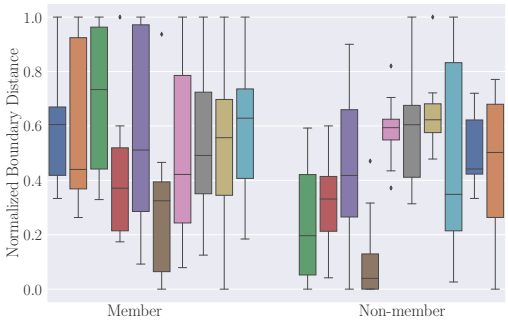


Figure 7: Normalized boundary distance of 10 members and 10 non-members in 100 random directions from VGG-16 trained on CIFAR10.

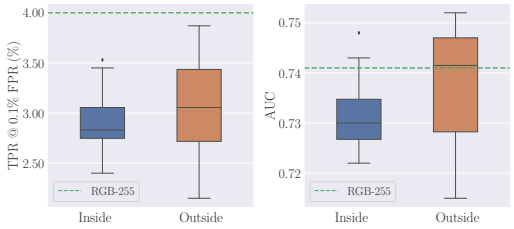


Figure 8: Performance of `DHAttack` with different fixed points on MobileNetV2 trained on CIFAR10.

ence, relScore mitigates the impact of sample diversity and provides better distinguishability, especially in the low FPR region. This indicates that the attack targeting a small portion of the dataset is very reliable.

5.3 Ablation Study

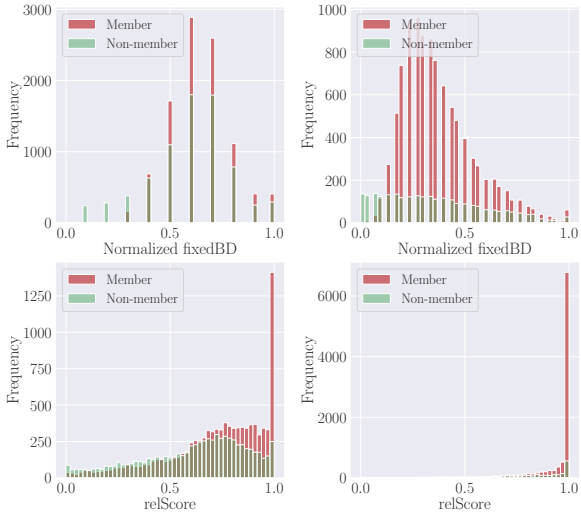
In this section, we perform ablation studies to investigate the influence of several important factors.

Number of Shadow Models. The Gaussian distribution of fixedBD values from shadow models is a key concept in our `DHAttack`, representing the non-member state of a target sample. However, the accuracy of this distribution, and consequently the attack performance, depends on the number of shadow models used. The relationship between them is presented in Figure 12 (See more results in Appendix Figure 20). We can see that the attack performance is significantly improved as the number of shadow models increases, especially for TPR @ 0.1% FPR. This is due to the fact that more shadow models can output more fixedBD values for a target sample, which leads to a more accurate Gaussian distribution. In addition, we also observe that after the number of models reaches a certain magnitude, e.g., between 128 and 256, the improvement of the attack performance is no longer significant, and for some cases, it even decreases slightly. For example, for ResNet-56 trained on CIFAR10, TPR @ 0.1% FPR with 192 shadow models is 2.38, which is lower than 2.64 with 128 shadow models, as shown in Figure 12. The improvement in AUC also becomes minimal between 64 and 256 shadow models. Therefore, the adversary can choose the proper number of shadow models according to their computational resources.

Size of Reference Dataset. The size of the reference dataset

(a) CIFAR10 (b) CIFAR100

Figure 9: Visualization of distributions for different datasets using t-SNE.



(a) CIFAR10, VGG-16 (b) CIFAR100, ResNet-56

Figure 10: The distributions of fixedBD and relScore values obtained from different models and datasets. Note that we exclude non-member samples that are easily identifiable (those with fixedBDs below 0.005), aiming to highlight its capability to distinguish between members and non-members with similar boundary distances.

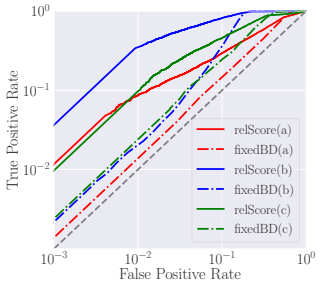


Figure 11: Log-scale ROC Curves for fixedBD and relScore. Scenarios: (a) VGG-16 with CIFAR10, (b) ResNet-56 with CIFAR100, and (c) MobileNetV2 with CINIC10.

is an important factor that affects how much a shadow model learns from the target model. Table 4 shows that an increase

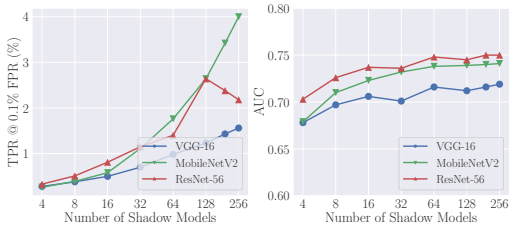


Figure 12: Attack performance under the effect of the number of shadow models. The dataset is CIFAR10.

Table 4: The impact of reference dataset size for MobileNetV2 trained on CIFAR10.

	Reference dataset size					
	1k	5k	10k	20k	30k	40k
TPR @ 0.1% FPR (%)	0.25	0.75	1.74	4.00	3.16	2.70
AUC	0.715	0.733	0.740	0.741	0.730	0.751

Table 5: The impact of the overfitting level of the target model. The experiments are conducted on MobileNetV2 trained on CIFAR10.

	Training dataset size				
	30k	25k	20k	15k	10k
Overfitting level	0.168	0.201	0.217	0.251	0.319
TPR @ 0.1% FPR (%)	0.96	1.32	1.14	1.41	4.00
AUC	0.631	0.656	0.658	0.697	0.741

in the size of the reference dataset does improve attack performance, as the shadow models learn better about the target model (See more results in Appendix Table 12). However, continually increasing the set size after it reaches a certain number does not lead to greater attack gains, but rather decreases them. For instance, `DHAttack` with 20k reference samples achieves 4.00 in terms of TPR @ 0.1% FPR, while it drops to 3.16 when using 30k reference samples. We attribute this to the fact that leveraging too many samples to train the shadow models results in a difference in the degree of generalization from the target model, which in turn can lead to a failure in the simulation of the target model. Upon experimentation, we find that a reference dataset with about 2 times the size of the training set is the most appropriate.

Overfitting Level of the Target Model. It is well known that overfitting of the target model is one of the main causes of membership privacy leakage [34, 37, 39, 46]. Therefore, we experimentally demonstrate the impact of target models with different overfitting levels on the performance of `DHAttack`. Following [23, 27, 34], we quantify the overfitting level L of a model as the difference between its training accuracy (Acc_{train}) and testing accuracy (Acc_{test}): $L = Acc_{train} - Acc_{test}$. Specifically, we manipulate L by controlling the size of the training set. Table 5 shows that the higher the degree of overfitting, the more effective the attack is. For example, the AUC

Table 6: Attack performance of DHAttack and baselines in two low-overfitting scenarios: overfitting levels 0.08 and 0.07.

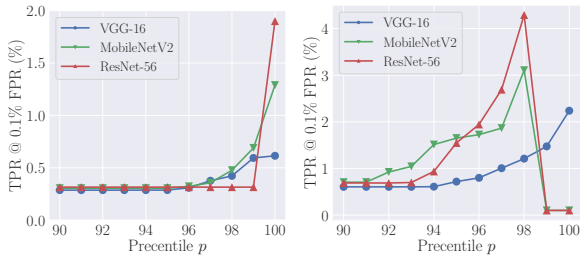
MIA method	MobileNetV2 (0.08)		ResNet-56 (0.07)	
	TPR @ 0.1% FPR (%)	AUC	TPR @ 0.1% FPR (%)	AUC
NRA	0.14	0.548	0.11	0.542
UBA	0.12	0.568	0.11	0.561
SBA	0.13	0.571	0.12	0.563
TrajectoryMIA	0.14	0.548	0.0	0.557
YOQO	0.12	0.565	0.11	0.558
DHAttack	0.68	0.573	0.16	0.560

of DHAttack is 0.741 when the training set size of the target model is 10k, while the AUC is 0.631 when the size is 30k. This is because the degree of overfitting is reduced from 0.319 to 0.168.

Furthermore, we introduce two low-overfitting scenarios of GTSRB [40] with overfitting levels of 0.08 and 0.07, respectively. We select GTSRB as it is a benchmark dataset used in previous MIAs [23, 26, 27] and, more importantly, allows for training models with very low overfitting levels. As shown in Table 6, although all methods show reduced performance in these scenarios as compared to Table 3, DHAttack significantly outperforms the baselines in terms of TPR @ 0.1% FPR, achieving 0.68 on MobileNetV2, against the highest baseline score of only 0.14, and matches them on the AUC metric. We attribute the decreased attack performance across all methods to reduced overfitting, which lessens the boundary distance differences between members and non-members, both for shortestBD and fixedBD. By calculating relScore using non-member state as a reference, DHAttack notably enhances distinguishability, especially improving the TPR @ low FPR.

Alternative fixedBD Measurements. In addition to perturbing target samples toward "**RGB-255**", we explore several alternative methods: (1) **Blurriness**, which gradually blurs the image by averaging each pixel with its neighbors, (2) **Resize**, which reduces the image dimensions, filling exposed pixels with zeros, and (3) **Rotation**, which rotates the image up to 180 degrees. We emphasize these methods are still integrated into the DHAttack framework, with modifications limited to the fixedBD computation. Accordingly, their fixedBDs are determined by the number of operations required until the ith operation does not alter the predicted label, but the $(i+1)$th operation does, which actually aligns with the method described earlier. Besides, we introduce another alternative method using "**RGB-0**" (a fully black image) as the fixed sample. As shown in Table 7, both "RGB-255" and "RGB-0" generally outperform other methods, with Blurriness also performing well. This indicates that employing a constant perturbation pattern to measure boundary distance is effective. Furthermore, selecting an outlier, such as "RGB-255" or "RGB-0", ensures that most samples cross the decision boundary, resulting in optimal attack performance.

Threshold Choosing. As discussed in Section 3.2, we propose a simple method for selecting the relScore threshold, using synthetic samples as non-members. Concretely, we uniformly sample 200 synthetic samples from the RGB range of 0-255, which are non-members, as they are unseen by the target model. We calculate and sort their relScore values, then select different top p percentiles as thresholds for membership inference. As shown in Figure 13, the optimal threshold for attack performance can be found between $p = 95\%$ and $p = 100\%$. Moreover, we observe that the TPR @ 0.1% FPR corresponding to the optimal threshold surpasses that of all other baselines. Thus, we can easily identify an effective threshold for DHAttack.



(a) CIFAR10 (b) CIFAR100

Figure 13: The relation between threshold and attack performance. The x-axis represents the top p percentile of relScore values and the y-axis represents TPR @ 0.1% FPR.

5.4 Practical Investigation

In this section, we will relax the assumptions about the attacker's knowledge and attempt to simplify some complex steps to make DHAttack more practical.

Relax Assumptions. To begin with, we relax the first assumption about the adversary's knowledge from Section 2.2, which assumes that the adversary possesses an auxiliary dataset D^r that is from the same distribution as the target model's training set D^t. Here, we use CIFAR10 as the training and testing set for the target model, while the adversary holds an auxiliary dataset from the ImageNet portion of CINIC10 (denoted as $D^t \neq D^r$). Figure 14 shows that the attack performance of DHAttack decreases when the distribution of the dataset varies. We attribute this to the fact that differences in data distribution can lead to deviations in the simulation of the target model's prediction behavior by the shadow models, thereby reducing the accuracy of the relScore calculation. However, we note that DHAttack ($D^t \neq D^r$) still outperforms baselines with $D^t = D^r$. For example, with $D^t \neq D^r$, DHAttack achieves a TPR of 0.86% at 0.1% FPR against VGG-16 (see Figure 14). In contrast, when $D^t = D^r$ (i.e., the adversary has stronger training knowledge), other baselines achieve at most 0.34% (see Table 3).

Second, we proceed to relax another assumption that the adversary has the knowledge of the architecture and hyper-

Table 7: Attack performance of DHAttack on CIFAR10 using alternative measurement methods for fixedBD. We highlight the top 2 performances of each metric in bold.

Different Methods	TPR @ 0.1% FPR (%)			AUC		
	VGG-16	ResNet-56	MobileNetV2	VGG-16	ResNet-56	MobileNetV2
Blurriness	**1.69**	0.90	**3.30**	0.642	0.597	0.653
Rotation	0.87	0.18	1.22	0.653	0.550	0.575
Resize	0.08	0.08	0.30	0.498	0.500	0.390
RGB-0	1.22	**2.53**	**4.04**	**0.719**	**0.753**	**0.756**
RGB-255	**1.56**	**2.58**	2.93	**0.719**	**0.752**	**0.750**

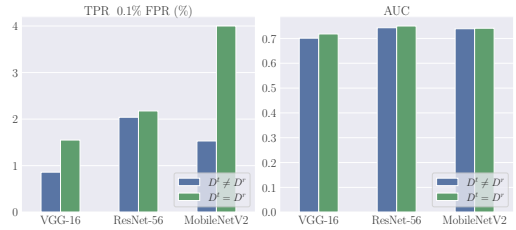


Figure 14: Attack performance of DHAttack when the adversary uses the same distribution as the target model's training set ($D^t = D^r$) versus different distributions ($D^t \neq D^r$).

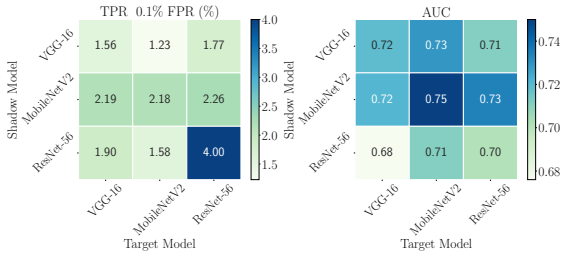


Figure 15: Attack performance of DHAttack using different model architectures for training shadow models. Both target and shadow models are trained on CIFAR10.

parameters of the target model. Figure 15 shows the attack performance when the adversary adopts different architectures and hyperparameters with the target model to train the shadow models. We can observe that the performance of DHAttack is optimal at most cases when using the same architecture and hyperparameters of the target model (i.e., along the diagonal). This can be attribute to the fact that the same model architecture and hyperparameters allow the shadow model to simulate the predicted behavior of the target model more accurately. Fortunately, the attack performance degradation due to different architectures is extremely small, as depicted in Figure 15. DHAttack with different model architectures outperforms baselines using the same architecture. For example, with MobileNetV2 as the target and VGG-16 as the shadow model, DHAttack achieves a TPR of 1.23% at an FPR of 0.1% (see Figure 15), significantly surpassing the 0.18% upper bound of baselines using MobileNetV2 for both models (see Table 3).

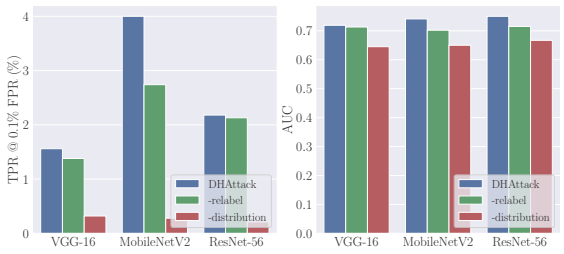


Figure 16: Impact of reference data relabeling and non-member state construction. Models are trained on CIFAR10.

Overall, even when relaxing the assumptions outlined in Section 2.2, our DHAttack consistently outperforms baselines that do not relax these assumptions. This superior performance can be attributed to the Reference Data Relabeling process in DHAttack, which, similar to knowledge distillation, transfers knowledge from the target model to the shadow models, even when their architectures and training data distributions differ.

Further Simplification of DHAttack. Here, we recall the entire DHAttack process in the hope of further simplifying our attack, including reducing the query times to the target model, as well as reducing the local computation.

First, we find that reference data relabeling launches a certain initial query to the target model in order to enable the shadow models to better learn the knowledge of the target model. Although this step is no longer executed in subsequent attacks, we still try to see if we can optimize away it. Figure 16 shows that when we train the local shadow models directly using the reference dataset without relabeling, denoted as "-relabel," its performance is slightly degraded compared to the whole DHAttack. For example, for VGG-16 trained on CIFAR10, TPR @ 0.1% FPR is reduced from 1.56 to 1.38. Therefore, this relabeling operation can be omitted when the adversary needs to further reduce the disturbance to the target model.

Second, we find that non-member state construction needs to train a large number of local shadow models, which is a computational burden on the adversary. Since our primary objective is to reduce the impact of sample diversity on fixedBD, we use the difference between the actual fixedBD from the target model and a single fixedBD from a shadow model, rather

Table 8: Performance of MIAs on Purchase and News.

MIA method	Purchase		News	
	TPR @ 0.1% FPR (%)	AUC	TPR @ 0.1% FPR (%)	AUC
NRA	0.21	0.715	0.20	0.592
UBA	0.15	0.517	0.20	0.531
SBA	0.21	**0.738**	0.32	**0.782**
DHAttack	**0.34**	0.727	**0.33**	0.761

than the relScore, as the membership signal. This approach is referred to as the "-distribution." Figure 16 shows that such a simplification significantly reduces the effectiveness of the attack, e.g., for CIFAR10 trained on VGG-16, TPR @ 0.1% FPR is reduced from 1.56 to 0.32. This indicates that the randomness in the model training algorithm significantly affects the fixedBD values from shadow models and suggests that using a Gaussian distribution to model these values is effective. However, we also find that 0.32 TPR @ 0.1% FPR is still higher than most baselines. Therefore, we recommend that when the attacker's computational resources are severely limited, this simplified approach can still yield reliable attack results to some extent.

6 Discussion

In this section, we evaluate the performance of `DHAttack` on non-image datasets and test its robustness on several existing defenses. Then, we further discuss the limitations.

6.1 Beyond Images

We compared `DHAttack` with baselines on Purchase and News, two non-image datasets commonly used in existing works [27, 34, 37, 43]. Since YOQO is computationally intensive and performs lower on non-image datasets than the baseline SBA (as demonstrated by themselves [43]), we do not put it into the performance comparison on non-image datasets here.

Specifically, the definitions of the shortestBD for UBA and SBA, as well as the fixedBD for `DHAttack`, are provided in Section 2.3. The sole difference for non-image data lies in the fixed sample selection. Instead of using "RGB-255" as in image datasets, we assign the maximum value across all dimensions to the sample as the fixed point. Table 8 illustrate the results on non-image datasets. We observe that `DHAttack` achieves the best TPR @ 0.1% FPR, but its AUC is suboptimal. For instance, the TPR @ 0.1% FPR of `DHAttack` is 0.34, while that of baselines is at most 0.21. However, compared to image datasets, our `DHAttack` performs less effectively on non-image data. We attribute this to the fact that it is harder to find an appropriate fixed point of non-image datasets than that of image datasets (such as an image with all RGB values set to 255). As we known, image samples take continuous values for each dimension, whereas non-images do not, e.g., the samples in Purchase take 0 or 1 for each dimension, representing whether or not a product is purchased. Moreover, neighboring dimensions in image samples are correlated, while non-image samples lack such strong correlations, e.g., each dimension in News represents a word in the vocabulary, resulting in weaker correlations between the neighboring dimensions. Overall, for image datasets, our fixedBD calculation method allows target samples to consistently and steadily move towards the fixed point and cross the decision boundary. However, this is difficult to achieve with non-image datasets.

6.2 Evaluation of Robustness

We consider several popular defenses to evaluate the robustness of `DHAttack`.

MixupMMD. MixupMMD is a generalization enhancement technique proposed by Li et al. [25]. It adds an extra regularization to the target model's training process to narrow the generalization gap between the training and testing accuracy. As previously discussed, the overfitting level is one of the main causes of membership privacy leakage. Therefore, MixupMMD does reduce the performance of all attaks, as shown in Table 9. However, `DHAttack` still achieves the best performance on the well-generalized model. For instance, TPR @ 0.1% FPR of `DHAttack` is 0.22, while that of other baselines is only 0.12.

DP-SGD. DP-SGD [3] is a general privacy-preserving method for machine learning, which adds Gaussian noise to the gradients during the training process of the target model. Following [7, 23, 27], we fix the hyperparameters $\delta = 1e\text{-}5$ and $C = 1$, while varying $\sigma = 0$, 0.2, 0.5, and 1 to control the privacy budget ε. Here, $\sigma = 0$ indicates no noise is added, with only gradient clipping (i.e., $C = 1$). A smaller ε indicates stronger defense. Table 10 shows that the performance of our attack diminishes as ε decreases. However, the accuracy of the target model also significantly decreases when stronger defense effects are applied. To provide an acceptable trade-off between the utility and privacy of the target model, we adopt $\sigma = 0.2$ of DP-SGD for further comparison. As depicted in Table 9, DP-SGD effectively mitigates the privacy leakage of the target model in the label-only scenario. However, our `DHAttack` still outperforms all baselines. For example, TPR @ 0.1% FPR of `DHAttack` is 0.20, while that of other baselines is at most 0.11.

LDL. LDL is a lightweight defense method proposed by Rajabi et al. [32], which is designed for label-only settings. Since it outputs the class corresponding to the mean of the posteriors of the target sample's neighbor samples within a high-dimensional sphere, instead of the original result, it effectively destroys the attacker's measurement of the decision boundary. Following [32], we attempt to set σ^2 to 0.02, 0.04, and 0.06, and finally fix $\sigma^2 = 0.06$ in our experiments, which achieves the best trade-off between the target model's utility and privacy. As shown in Table 9, TPR @ 0.1% FPR of

Table 9: Performance of different attacks against VGG-16 trained on CIFAR10 with different defenses.

MIA method	TPR @ 0.1% FPR (%)				AUC			
	No defense	MixupMMD	DP-SGD	LDL	No defense	MixupMMD	DP-SGD	LDL
NRA	0.18	0.12	0.10	0.14	0.700	0.548	0.511	0.615
UBA	0.19	0.12	0.11	0.15	0.721	0.545	0.501	0.608
SBA	0.19	0.12	0.11	0.14	0.722	0.546	0.520	**0.638**
DHAttack	1.56	**0.22**	**0.20**	**0.54**	0.719	**0.564**	**0.538**	0.620

Table 10: Performance of `DHAttack` against VGG-16 trained on CIFAR10 with DP-SGD ($\delta = 1e-5$ and $C = 1$).

	DP-SGD		Accuracy of the target model	TPR @ 0.1% FPR (%)	AUC
	σ	ε			
No defense	-	-	0.756	1.56	0.718
DP-SGD	0	∞	0.575	0.83	0.706
	0.2	1523	0.581	0.20	0.538
	0.5	43	0.482	0.09	0.515
	1	6	0.377	0.12	0.502

`DHAttack` is three times higher than that of other baselines.

The reason is that `DHAttack` uses the location where the prediction of noisy samples changes as the boundary distance (see Figure 4). As a result, LDL can only cause the boundary distance measured by `DHAttack` to be inaccurate by a few scale units. However, SBA and UBA use the shortest distance between the target sample and its adversarial example, which can be perturbed by LDL with a significant error.

6.3 Limitations

`DHAttack` does not work well against models with non-image tasks, as discussed in Section 6.1. The reason is that our method of measuring the boundary distance is not suitable for non-image samples. We leave the in-depth exploration of more effective boundary distance measurements for non-image datasets as future work.

7 Related Work

7.1 Membership Inference Attacks

Most existing MIAs assume that members receive higher confidence outputs from the target model than non-members. In black-box settings, Shokri et al. [37] and Salem et al. [34] introduced shadow training to mimic the target model's behavior and trained an attack model based on shadow models' output posteriors. Song et al. [39] and Yeom et al. [46] later inferred membership directly using metrics like loss, which typically favor members. Additionally, Hui et al. [20] proposed an attack that uses differential comparison to infer membership based on the model's output probability distributions. Furthermore, to address high false positives, researchers [6, 7, 33, 42, 45] calibrated the loss metric using each sample's hardness threshold. Liu et al. [27] and Li et al. [23] further enhanced attacks by employing knowledge distillation to extract additional information from the target model's training.

However, in label-only settings, traditional methods using output posteriors are not applicable. Li et al. [26] and Choquette-Choo et al. [11] introduced boundary distance, the distance of a sample from the target model's decision boundary, as the membership signal. They found that members typically have larger boundary distances than non-members. Later, [44, 48] improved these methods by adjusting the directions of adversarial perturbations. Liu et al. [27] presented TrajectoryMIA (label-only), which uses additional membership signals generated during the target model's training, alongside boundary distance. These methods often require over 1,000 queries to determine each sample's boundary distance. Wu et al. [43] addressed this by introducing the concept of improvement area, reducing the required queries to just one while maintaining comparable performance. Additionally, Chaudhari et al. [8] proposed a novel attack method using data poisoning, achieving high performance in the low False Positive Rate (FPR) regime.

7.2 Defenses Against MIAs

One simple but effective defence method is reducing the overfitting level of target model. Dropout, L2 regularization and label smoothing have been used by [28, 34, 37]. Additionally, Li et al. [25] elaborate a novel method MixupMMD to narrow the target model's generalization gap for mitigating MIAs. Another strategy is to perturb the target model's output, such as MemGuard [21]. For the label-only settings, Rajabi et al. [32] proposed LDL, a light weight defense against label-only MIAs, which constructs a high-dimensional sphere around the target sample, and output the same decision for all query samples in the sphere. Therefore, the adversary can not obtain the accurate boundary distance. Finally, DP-SGD [3] is a traditional privacy-preserving method, which is also suitable for the label-only settings. It adds differential privacy [15] for the stochastic gradient descent algorithm, and thus obscure the differences between members and non-members.

8 Conclusion

In this work, we propose a higher performance and higher stealth label-only MIA, called `DHAttack`, which utilizes a new

sample-dependent membership signal in label-only scenarios. This signal represents the probability that a sample's actual boundary distance on the target model exceeds the boundary distances drawn from its non-member state distribution. Our extensive experiments demonstrate that `DHAttack` achieves the highest AUC scores, especially TPR @ low FPR, with few queries on the target model. Furthermore, we conduct experiments to investigate the reasons behind our high attack performance, and analyze some important factors affecting `DHAttack`. Finally, we perform `DHAttack` on non-image datasets and evaluate its robustness on several existing defenses. In the future, we aim to investigate how to improve the measurement of a sample's boundary distance for non-image tasks and extend our attack to more scenarios.

9 Acknowledgements

We thank all anonymous reviewers for their constructive comments. This work is supported by National Key R&D Program of China (2022YFB4501500, 2022YFB4501503), the European Health and Digital Executive Agency (HADEA) within the project "Understanding the individual host response against Hepatitis D Virus to develop a personalized approach for the management of hepatitis D" (DSolve, grant agreement number 101057917) and the BMBF with the project "Repräsentative, synthetische Gesundheitsdaten mit starken Privatsphärengarantien" (PriSyn, 16KISAO29K).

Ethics Considerations

This paper presents a membership inference attack that poses a significant risk to the privacy of training data in machine learning models in label-only settings. To mitigate potential misuse of the proposed attack, we conduct all experiments exclusively on public datasets and widely used model architectures. These datasets are not related to any privacy concerns. Despite the risk of privacy leaks, we believe that raising awareness about the membership leakage problem is crucial, as it encourages model holders and the research community to develop stronger defense mechanisms. Furthermore, our method can be utilized to more thoroughly assess the privacy leakage risks within a model, which may, in turn, help enhance the privacy safeguards of existing models.

Open Science

For the open science policy, all datasets used in our research, including CIFAR10, CIFAR100, CINIC10, GTSRB, Purchase and News, are open source and can be accessed from their respective official websites. Additionally, the source code for our attack `DHAttack` will be made available after paper acceptance and before the camera-ready submission deadline.

References

[1] https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data.

[2] http://people.csail.mit.edu/jrennie/20Newsgroups.

[3] 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.

[4] Ruibin Bai, Xinan Chen, Zhi-Long Chen, Tianxiang Cui, Shuhui Gong, Wentao He, Xiaoping Jiang, Huan Jin, Jiahuan Jin, Graham Kendall, et al. Analytics and machine learning in vehicle routing research. *International Journal of Production Research*, 61(1):4–30, 2023.

[5] Batuhan Balci, Dan Saadati, and Dan Shiferaw. Handwritten text recognition using deep learning. *CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University, Course Project Report, Spring*, pages 752–759, 2017.

[6] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z. Wu. Scalable membership inference attacks via quantile regression. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 314–330. Curran Associates, Inc., 2023.

[7] 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.

[8] Harsh Chaudhari, Giorgio Severi, Alina Oprea, and Jonathan Ullman. Chameleon: Increasing label-only membership leakage with adaptive poisoning. *arXiv preprint arXiv:2310.03838*, 2023.

[9] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[10] 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*, CCS '21, page 896–911. Association for Computing Machinery, 2021.

[11] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR, 2021.

[12] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[13] 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.

[14] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004.

[15] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

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

[17] Yu He, Boheng Li, Yao Wang, Mengda Yang, Juan Wang, Hongxin Hu, and Xingyu Zhao. Is difficulty calibration all we need? towards more practical membership inference attacks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1226–1240, 2024.

[18] Sorami Hisamoto, Matt Post, and Kevin Duh. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63, 2020.

[19] Jen-Hao Hsiao and Li-Jia Li. On visual similarity based interactive product recommendation for online shopping. In *2014 IEEE international conference on image processing (ICIP)*, pages 3038–3041. IEEE, 2014.

[20] 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.

[21] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 259–274, 2019.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[23] Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. Seqmia: Sequential-metric based membership inference attack. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3496–3510, 2024.

[24] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Qeba: Query-efficient boundary-based black-box attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1221–1230, 2020.

[25] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, CODASPY '21, page 5–16. Association for Computing Machinery, 2021.

[26] Zheng Li and Yang Zhang. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 880–895. Association for Computing Machinery, 2021.

[27] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 2085–2098. Association for Computing Machinery, 2022.

[28] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4525–4542. USENIX Association, 2022.

[29] 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.

[30] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. Dropout improves recurrent neural networks for handwriting recognition. In *2014 14th*

international conference on frontiers in handwriting recognition, pages 285–290. IEEE, 2014.

[31] Réjean Plamondon and Sargur N Srihari. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):63–84, 2000.

[32] Arezoo Rajabi, Dinuka Sahabandu, Luyao Niu, Bhaskar Ramasubramanian, and Radha Poovendran. Ldl: A defense for label-based membership inference attacks. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, pages 95–108, 2023.

[33] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR, 2019.

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

[35] 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.

[36] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.

[37] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE Computer Society, 2017.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] 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. USENIX Association, 2021.

[40] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

[41] Abd-Elhamid Taha and Najah AbuAli. Route planning considerations for autonomous vehicles. *IEEE Communications Magazine*, 56(10):78–84, 2018.

[42] Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022.

[43] YUTONG WU, Han Qiu, Shangwei Guo, Jiwei Li, and Tianwei Zhang. You only query once: An efficient label-only membership inference attack. In *The Twelfth International Conference on Learning Representations*, 2023.

[44] JiaCheng Xu and ChengXiang Tan. Membership inference attack with relative decision boundary distance. *arXiv preprint arXiv:2306.04109*, 2023.

[45] 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*, CCS '22, page 3093–3106. Association for Computing Machinery, 2022.

[46] 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.

[47] 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*, CCS '21, page 864–879. Association for Computing Machinery, 2021.

[48] Zhaoxi Zhang, Leo Yu Zhang, Xufei Zheng, Bilal Hussain Abbasi, and Shengshan Hu. Evaluating membership inference through adversarial robustness. *The Computer Journal*, 65(11):2969–2978, 2022.

A Dataset Description

CIFAR10/CIFAR100. CIFAR10 and CIFAR100 are popular benchmark datasets for evaluating image recognition algorithms. Each of them includes 60,000 images of size

$3 \times 32 \times 32$. The difference is that CIFAR10 has 10 classes, while CIFAR100 has 100 classes.

CINIC10. CINIC10 contains 270,000 images within the same classes as CIFAR10 (210,000 images from ImageNet [13] and 60,000 images from CIFAR10). In this paper, we just sample 60,000 images from CINIC10 for our experiments, and thus, most images are from ImageNet.

Purchase. Purchase is a dataset extracted from Kaggle's "acquire valued shopper" challenge, which contains 197,324 samples of 600 dimensions. Each sample is a purchase record, and each dimension in the record indicates whether one product is purchased. Following [37], we cluster 197,324 records into 100 classes to generate their ground-truth labels.

News. News (20 Newsgroups) is a common benchmark dataset for text classification, which contains 20,000 documents within 20 classes. Following [34], we build the TF-IDF form of each document, which is a vector of 134,410 dimensions.

GTSRB. GTSRB is a dataset for traffic sign recognition, consisting of over 50,000 images across 43 different classes of traffic signs.

Table 11: Data splits for our evaluation.

Dataset	$\mathcal{D}^t_{train}$	$\mathcal{D}^t_{test}$	$\mathcal{D}^s_{train}$	$\mathcal{D}^s_{test}$	$\mathcal{D}^r$
CIFAR10	10000	10000	10000	10000	20000
CIFAR100	10000	10000	10000	10000	20000
CINIC10	10000	10000	10000	10000	20000
Purchase	20000	20000	20000	20000	40000
News	3000	3000	3000	3000	6000
GTSRB	1500	1500	1500	1500	45000

B Algorithms

Algorithm 1: DHAttack

Input: Reference Data D^r, target sample x and its label y, target model f_T, number of shadow models n, fixed sample x_{fixed}, number of queries K, threshold τ.

Output: Relative membership score *relScore*.

// Relabel each sample in D^r.
for $x_i \in D^r$ **do**
  $y_i \leftarrow f_T(x_i)$;
end
// Train n shadow models.
for $i \leftarrow 1$ **to** n **do**
  Train θ_i with D^r;
end
// Construct Non-member state of x with Algorithm 2.
for $i \leftarrow 1$ **to** n **do**
  $d_i \leftarrow \text{fixedBD}(x, y, \theta_i, x_{\text{fixed}}, K)$;
end
$\mu \leftarrow \frac{1}{n}\sum_{i=1}^{n} d_i$;
$\sigma^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n}(d_i - \mu)^2$;
$G \leftarrow \mathcal{N}(\mu, \sigma^2)$;
// Membership inference
$d \leftarrow \text{fixedBD}(x, y, f_T, x_{\text{fixed}}, K)$;
$relScore \leftarrow \text{CDF}(d, G)$;
return $\mathbb{I}[relScore > \tau]$;

C Additional Experimental Results

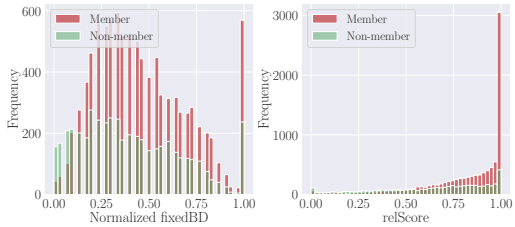


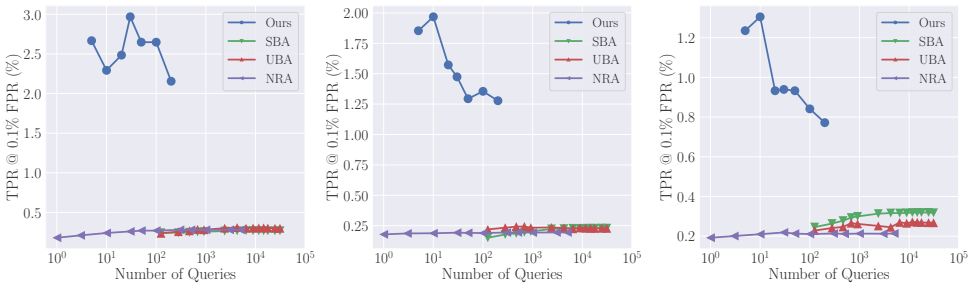
Figure 17: The distributions of fixedBD and relScore values obtained from MobileNetV2 trained on CINIC10.

Algorithm 2: FixedBD

Input: Target sample x and its label y, model f, fixed sample x_{fixed}, and the number of segments (i.e., the maximum number of queries) K.

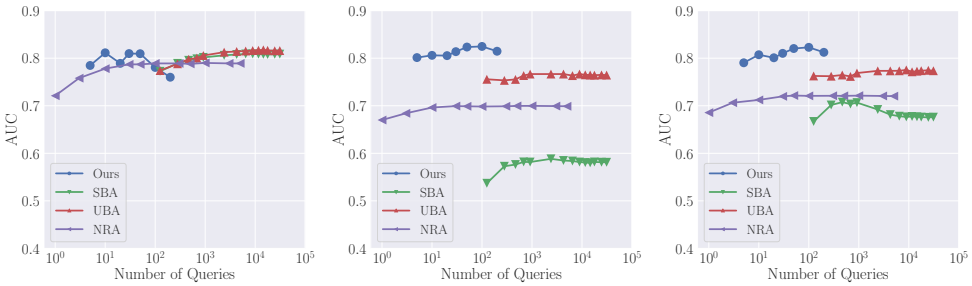
Output: FixedBD d.

$x_{\text{diff}} \leftarrow x_{\text{fixed}} - x$;
// Query model f up to K times to calculate fixedBD d.
$d \leftarrow K$;
for $i \leftarrow 0$ **to** K **do**
  $x_{\text{masked}} \leftarrow \frac{i}{K} \cdot x_{\text{diff}} + x$;
  $\hat{y} \leftarrow f(x_{\text{masked}})$;
  if $\hat{y} \neq y$ **then**
    $d \leftarrow i$;
    break;
  end
end
return d;



(a) VGG-16 (b) ResNet-56 (c) MobileNetV2

Figure 18: TPR @ 0.1% FPR under different numbers of queries for attacks on three model architectures (CINIC10 Dataset).



(a) VGG-16 (b) ResNet-56 (c) MobileNetV2

Figure 19: AUC under different number of queries for attacks on three model architectures (CINIC10 Dataset).

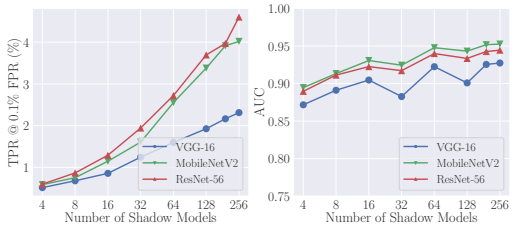


Figure 20: Attack performance under the effect of the number of shadow models: experiments on different models trained on CIFAR100.

Table 12: The impact of reference dataset size for MobileNetV2 trained on CIFAR100.

	Reference dataset size					
	1k	5k	10k	20k	30k	40k
TPR @ 0.1% FPR (%)	0.67	1.27	2.80	4.02	5.74	5.90
AUC	0.921	0.940	0.950	0.954	0.954	0.955