SeqMIA: Sequential-Metric Based Membership Inference Attack

Hao Li*
Institute of Software, Chinese Academy of Sciences
Beijing, China

Zheng Li*
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Siyuan Wu
Institute of Software, Chinese Academy of Sciences
Beijing, China

Chengrui Hu
Institute of Software, Chinese Academy of Sciences
Beijing, China

Yutong Ye
Institute of Software, Chinese Academy of Sciences
Zhongguancun Laboratory
Beijing, China

Min Zhang†
Institute of Software, Chinese Academy of Sciences
Beijing, China

Dengguo Feng
Institute of Software, Chinese Academy of Sciences
Beijing, China

Yang Zhang
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Abstract

Most existing membership inference attacks (MIAs) utilize metrics (e.g., loss) calculated on the model's final state, while recent advanced attacks leverage metrics computed at various stages, including both intermediate and final stages, throughout the model training. Nevertheless, these attacks often process multiple intermediate states of the metric independently, ignoring their time-dependent patterns. Consequently, they struggle to effectively distinguish between members and non-members who exhibit similar metric values, particularly resulting in a high false-positive rate.

In this study, we delve deeper into the new membership signals in the black-box scenario. We identify a new, more integrated membership signal: *the Pattern of Metric Sequence*, derived from the various stages of model training. We contend that current signals provide only partial perspectives of this new signal: the new one encompasses both the model's multiple intermediate and final states, with a greater emphasis on temporal patterns among them. Building upon this signal, we introduce a novel attack method called Sequential-metric based Membership Inference Attack (`SeqMIA`). Specifically, we utilize knowledge distillation to obtain a set of distilled models representing various stages of the target model's training. We then assess multiple metrics on these distilled models in chronological order, creating *distilled metric sequence*. We finally integrate distilled multi-metric sequences as a sequential multiformat and employ an attention-based RNN attack model for inference. Empirical results show `SeqMIA` outperforms all baselines, especially can achieve an order of magnitude improvement in terms of TPR @ 0.1% FPR. Furthermore, we delve into the reasons why this signal contributes to `SeqMIA`'s high attack performance, and assess various defense mechanisms against `SeqMIA`.[1]

CCS Concepts

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

Keywords

Membership Inference, Metric Sequence, Knowledge Distillation

ACM Reference Format:
Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. 2024. SeqMIA: Sequential-Metric Based Membership Inference Attack. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3690335

1 Introduction

Machine learning (ML) has developed rapidly in the past decade. Unfortunately, existing studies [15, 16, 51] have shown that ML models can leak private information about their training set. Membership inference attacks (MIAs) [51] is one of the main privacy attacks that have attracted lots of researchers' concerns. It aims to infer whether a sample belongs to a model's training set, which in turn violates the privacy of the sample's owner. For example, if an ML model is trained on data collected from individuals with a certain disease, an adversary who knows that a victim's data belongs to the training data of the model can quickly infer the victim's health status.

Most existing studies [9, 24, 47, 51, 53, 69] employ the target model's output posteriors or some metric (e.g., loss) derived from them to launch their attacks. These attacks demonstrate effectiveness in average-case metrics such as balanced accuracy and

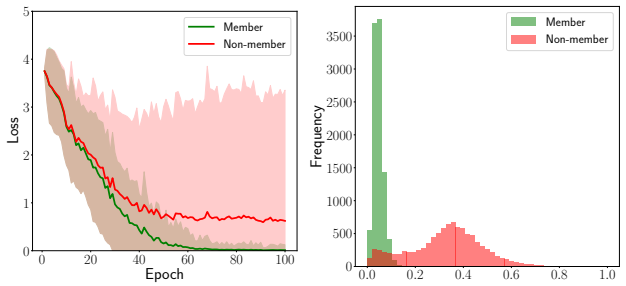
*The first two authors made equal contributions.
†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CCS '24, October 14–18, 2024, Salt Lake City, UT, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690335

[1]Our code is available at https://github.com/AIPAG/SeqMIA

Table 1: Various perspectives of the membership signal "Pattern of Metric Sequences." "✓" means this attack is based on this perspective and "-" indicates that it is not.

Attacks	Pattern of Metric Sequences (i.e., metric values overall training epochs)					
	Final State	Middle States	Time-Dependent Patterns			
			Fluctuation	Correlation	Decline Rate	Other Possible Implicit Patterns
[5, 6, 34, 45, 47, 51, 53, 57, 65]	✓	–	–	–	–	–
TrajectoryMIA [34]	✓	✓	–	–	–	–
SeqMIA	✓	✓	✓	✓	✓	✓



(a) Fluctuation Area of Loss **(b) Distribution of CLFA**

Figure 1: (a) the mean curves and fluctuation area of loss values for members and non-members during different training epochs; (b) the distribution of the cumulative loss fluctuation amplitude (CLFA) within 100 epochs.

ROC-AUC due to members typically exhibiting smaller losses compared to non-members. However, these attacks exhibit a high false-positive rate (FPR) when encountering both members and non-members with similar small losses. A high false positive rate means that an attacker will incorrectly identify non-member samples as members, thereby reducing the attack's effectiveness and reliability.

To tackle this issue, recent studies [6, 65] have employed sample-dependent thresholds to calibrate membership inference based on the target model, i.e., final model state at 100th (see Figure 1a). An alternative approach, known as TrajectoryMIA [34], introduces an additional membership signal, which is a collection of loss values gathered during the target model training process (i.e., 0∼100th epochs). The loss value set derived from various model states can reveal greater distinctions between members and non-members, even when they show similar low losses in the final model state. The findings of our experiments, however, indicate that these recent studies still face challenges in effectively distinguishing between members and non-members with similar sets of loss values, leading in particular to significantly higher false positive rates (FPR).

1.1 Our contributions

To overcome these limitations, in this work, we make an attempt to answer, "is it possible to explore a new membership signal that enhances the distinguishability between members and non-members, with a specific focus on reducing false-positive rate?"

Fortunately, we have discovered a new membership signal termed *the Pattern of Metric Sequence*, which is also derived from the various stages of model training. As Table 1 illustrates, we claim that the aforementioned signals provide only partial perspectives of this new signal: this new signal includes both the model's multiple intermediate and final states and focuses more on time-dependent patterns among them. To our knowledge, this signal has not been previously recognized or utilized in prior literature. Intuitively, we verify this signal from time-dependent views, such as fluctuation, correlation, and decline rate. We now illustrate the first two perspectives (see decline rate in Appendix B).

Fluctuation of Metric Sequences. We choose the most commonly used metric, loss, as our example. Figure 1a shows the sequence of loss values as the training progresses (denoted as loss sequence). Interestingly, we have further observed a new difference: *the fluctuation of loss sequence* between members and non-members also exhibits significant differences. More concretely, the loss sequence fluctuation of members tends to be smaller than that of non-members, especially around the 60th to 100th epoch. Besides expressing such fluctuation qualitatively, we further measure them quantitatively. Specifically, we compute the cumulative loss fluctuation amplitude (CLFA) for each sample by measuring the loss variation across *consecutive epochs*. We then count the frequency of the samples regarding their CLFA distribution. As depicted by Figure 1b, we observe members exhibit significantly smaller fluctuation of loss sequence compared to non-members. The results confirm that there exists a very clear difference in the pattern of loss sequence between members and non-members (see other metrics in Appendix Figure 13). Note that this observation is time-dependent and can only be observed in metric sequence as the training epoch progresses, unlike the *loss set* used in TrajectoryMIA, where shuffling the order does not affect the attack performance (see Section 5.2).

Correlation between Metric Sequences. Building upon the metric sequence, we delve into another new view: the correlation between two different metric sequences. The intuition is that two kinds of metric sequences (e.g., loss sequence and entropy sequence) of members tend to follow a similar trend compared to non-members, as the model is trained on members. Figure 2 presents the correlation coefficients among multiple sequences metrics. We observe that every pair of metric sequences for members shows correlation coefficients no smaller than those for non-members and, in most cases, even larger ones.

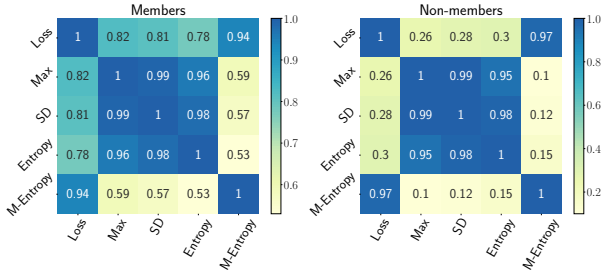


Figure 2: Absolute correlation coefficients among multiple metrics calculated from MLPs trained on Location.

SeqMIA. Building upon the pattern of metric sequences, we introduce a novel membership inference attack named SeqMIA (Sequential-metric based Membership Inference Attack). First, the adversary employs knowledge distillation to obtain a set of distilled models representing various stages of the target model's training. Then, the adversary assesses multiple metrics on these distilled models in chronological order, creating *distilled metric sequence*. The adversary integrates multiple distilled metric sequences into a sequential multiformat and utilizes a simple yet highly effective approach to handle this sequential data, namely employing an attention-based recurrent neural network (attention-based RNN) as the attack model for inference. This attention-based RNN can automatically excavate the aforementioned different patterns of metric sequence (even some more complex implicit patterns) without explicitly characterizing them in advance.

We conduct extensive experiments on 4 popular models using 7 benchmark datasets (4 image and 3 non-image datasets). Empirical results show that SeqMIA outperforms the baselines in nearly all cases. For example, when focusing on VGG-16 trained on CIFAR100, SeqMIA surpasses all baselines by more than an order of magnitude in terms of TPR @ 0.1% FPR. In addition, we conduct in-depth comparative analyses of metric non-sequences vs. metric sequences, and single vs. multiple metrics, revealing the reasons for the superior performance of SeqMIA. We also conduct ablation studies to analyze various factors on the attack performance. Finally, we demonstrate that SeqMIA performs better against several defenses compared to the baselines, especially at TPR @ 0.1% FPR. In general, our contributions can be summarized as follows:

- We introduce a novel membership signal termed *the Pattern of Metric Sequence*, which can more effectively capture the differences between members and non-members.
- We propose the Sequential-metric based Membership Inference Attack (called SeqMIA), which acquires sequential multi-metric from the target model's training process using knowledge distillation, then captures the membership signal via attention-based RNN attack model automatically.
- We extensively experiment and demonstrate that SeqMIA consistently outperforms all baselines, particularly in reducing the FPR by more than an order of magnitude.
- We conduct comprehensive analyses of the features of sequential membership signals, some key factors influencing attack performance, and various defenses against SeqMIA.

2 Preliminaries

2.1 Membership Inference Attack

Membership inference attack is one of the most popular privacy attacks against ML models. The goal of the membership inference attack is to determine whether a data sample is used to train a target model. We consider data samples as members if they are used to train the target model, otherwise, non-members. Formally, considering a data sample x, a trained ML model $\mathcal{M}$, and background knowledge of an adversary, denoted by $\mathcal{I}$, the membership inference attack $\mathcal{A}$ can be defined as the following:

$$\mathcal{A} : x, \mathcal{M}, \mathcal{I} \rightarrow \{0, 1\}.$$

Here, 0 means the data sample x is not a member of $\mathcal{M}$'s training dataset, and 1 otherwise. The attack model $\mathcal{A}$ is essentially a binary classifier.

2.2 Metrics for MIA

The success of existing membership inference is attributed to the inherent overfitting properties of ML models, i.e., models are more confident when faced with the data samples on which they are trained. This confidence is reflected in the model's output posterior, which results in several metrics that effectively differentiate between members and non-members. We are here to present a brief introduction below:

Loss. Loss, also known as the cost or objective function, measures how well an ML model's predictions match the ground truth for given data samples. The goal of the ML algorithm is to minimize this loss, as a lower loss indicates better performance of the model. Typically, members' losses are much lower than non-members' losses, and most existing works [6, 45, 57, 65, 66] leverage this discrepancy to mount their membership inferences.

Furthermore, loss trajectory is proposed by [34], which is a set of multiple losses from a model's training process, and is implemented as a vector. We here emphasize that the loss trajectory is not a sequential signal, due to the fact that there is no order between these loss values. If we swap the positions of losses in this vector, we will find that the attack performance of [34] is unaffected (see Section 5.2). Thus, we denoted it as *loss set* in the following sections in order to clearly indicate its essential features.

Max. Max refers to the maximum in the model's output posteriors, which is usually represented as a set of probabilities. To obtain single predicted class labels from these probabilities, one common approach is to take the class with the highest probability, i.e., maximum value. Similarly, the maximum value of members is usually greater than that of non-members, which has been used in [47, 53].

SD. Standard deviation is a measure of the dispersion of the model's output posterior from its mean. Members tend to have larger standard deviations than non-members because the model has more confidence in the predictions of the members, i.e., the probability of the correct class is greater, and the probability of the other class is much less. This metric has been used in [47].

Entropy. Entropy measures the uncertainty or randomness in a model's prediction. A low entropy indicates that the probability distribution is concentrated and the model is more certain about its predictions, while a high entropy indicates more uncertainty.

Similarly, the entropy value of members is lower than that of non-members, which has been used in [47, 51, 53]

M-Entropy. In contrast to entropy, which contains only information about the output posterior, modified entropy (M-Entropy) measures the model prediction uncertainty given the ground truth label. Thus, correct prediction with probability 1 leads to a modified entropy of 0, while incorrect prediction with probability 1 leads to a modified entropy of infinity. Also, the modified entropy of members is usually lower than that of non-members, and this metric is used in [53].

2.3 Knowledge Distillation

Knowledge distillation (KD) is a category of methods that transfer knowledge from large, complex models to smaller, more lightweight ones. The primary goal is to improve the performance of the smaller model while reducing resource consumption during deployment. The main idea is to use the soft information (i.e., the output posterior) of a larger teacher model as a supervised signal to train a smaller student model. This soft information contains more valuable knowledge than hard ground truth labels, leading to better generalization and efficiency of the student model.

Similar to [34], we use knowledge distillation to train a distilled model (student model) that is as close as possible to the target model (teacher model). In this work, we adopt the most widely-used KD framework proposed by Hinton et al. [23]. Concretely, we use a set of data (called distillation dataset) to query the teacher model and obtain its output posteriors, called soft labels. Then, when training the student model, soft labels are used to calculate the loss function in addition to the ground truth labels. The loss function can be expressed as follows:

$$L = \alpha L_{soft} + (1 - \alpha) L_{ground} \tag{1}$$

where L_{soft} is the Kullback-Leibler divergence loss between the soft labels and the student model's output posteriors, L_{ground} is the cross-entropy loss between the student model's output posteriors and the ground truth labels, and α is a weight coefficient.

Note that our goal is to simulate the target model training process and snapshot its intermediate version, rather than transferring knowledge from the larger model to the smaller one. Therefore, we employ the same model architecture as the target model to build the distilled model. Further, we set $\alpha = 1$, which means that the distilled model only mimics the target model's output posteriors regardless of the ground truth labels. For the sake of description, the intermediate distilled models obtained by distillation, are named as *snapshots* in this paper.

3 Attack Methodology

In this section, we present the attack methodology of `SeqMIA`. We start by introducing the threat model. Then, we describe the design intuition. Lastly, we present the detailed pipeline of `SeqMIA`.

3.1 Threat Model

In this paper, we focus on membership inference attacks in black-box scenarios, which means that the adversary can only access the output of the target model. Specifically, we only consider the case where the output is the predicted probability (posterior) rather than the predicted class label. Furthermore, we make two assumptions about the adversary's knowledge. First, the adversary holds a dataset D^a, which is from the same distribution as the target model's training dataset. Second, the adversary knows the architecture and hyperparameters of the target model. Such settings are following previous MIAs [6, 34, 38, 47, 51, 53, 65, 66]. Moreover, we further demonstrate in Section 5.3 that both of these assumptions can be relaxed.

3.2 Design Intuition

As aforementioned, we introduce a new membership signal termed *the Pattern of Metric Sequence*, which is also derived from the various stages of model training. This new signal includes both the model's multiple intermediate and final states but focuses more on time-dependent patterns among them. For example, members' metric sequences tend to demonstrate relatively smaller fluctuations compared to non-members. In addition, the correlations between different metric sequences of members are also much higher compared to non-members. Therefore, our general hypothesis is that simultaneous utilization of multiple metric sequences (serialized metric values) would yield significantly stronger membership signals compared to relying solely on a single metric or a non-serialized metric. Based on this insight, our first attack strategy is to construct "multi-metric sequences," which carry the pattern of metric sequences.

Furthermore, the previous study by Liu et al. [34] treats multiple losses from the various model states as a one-dimensional vector and directly feeds it into an MLP attack model for inference. However, the MLP model is primarily designed for independent input values and fails to capture the sequential or time-series information present in the input vector. This means that the MLP model may overlook important sequence-based signals in the input (see shuffling the vector's loss values in Section 5.2). In contrast, models specifically designed for time-series data, such as Recurrent Neural Networks (RNNs), are better able to capture the sequential information in the input vector, and thus can potentially excavate the sequence-based signals, e.g., fluctuations in the metric sequences as training progresses. Therefore, our second attack strategy involves using an attention-based RNN as the attack model to process the multiple metric sequences. This way, we can automatically uncover not only these explicit patterns but also more complex implicit patterns (see Section 5.2).

3.3 Attack Method

Based on the above, we propose a new membership inference attack, namely the Sequential-metric based Membership Inference Attack (`SeqMIA`). To execute `SeqMIA`, the adversary needs to acquire the multi-metric sequences from the training process of the target model. However, in this work, we consider the black-box scenario where the adversary can only access the final well-trained target model, i.e., the version at its final training epoch. To address this issue, similar to Liu et al. [34], the adversary leverages knowledge distillation on the target model to obtain its distilled model. This way, the adversary gains full control of the distillation process and can save the distilled models at different epochs. The attacker then

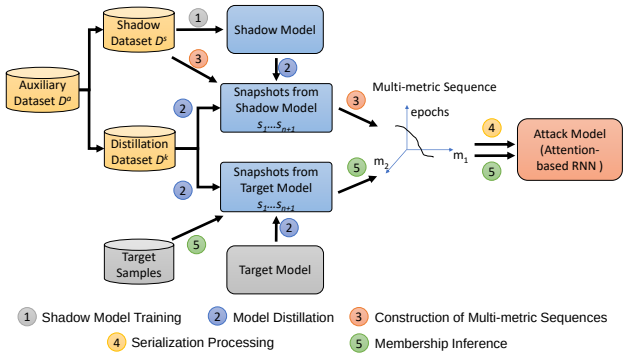


Figure 3: Overview of SeqMIA. Different from existing MIAs, SeqMIA focuses on the sequential membership information (multi-metric sequences) in a high-dimensional space.

evaluates different metrics of a given target sample on each intermediate distillation model to obtain its multi-metric sequence, called *distilled multi-metric sequence*. Finally, the attack model, functioning as a membership classifier, takes the distilled multi-metric sequence as input to infer the sample's membership status. The overview of SeqMIA is depicted in Figure 3, involving five stages: shadow model training, model distillation, construction of multi-metric sequences, serialization processing, and membership inference.

Shadow Model Training. As mentioned earlier, the adversary holds an auxiliary dataset D^a, which follows the same distribution as the target model's training dataset. The adversary first divides this auxiliary dataset D^a into two disjoint sets: the shadow dataset D^s and the distillation dataset D^k. The shadow dataset D^s is divided into two disjoint datasets, namely D^s_{train} and D^s_{test}. D^s_{train}, representing the members, is utilized to train the shadow model, which aims to emulate the behavior of the target model, while D^s_{test} represents the non-members. Given the assumption specified in Section 3.1, the adversary can train a shadow model with the same architecture and hyperparameters of the target model.

Model Distillation. The distillation dataset D^k is used to distill the target and shadow models, simulating their training process. For brevity, we refer to the target model and shadow model as the original models. Following the approach in Liu et al. [34], we query the two original models to obtain their output posteriors as soft labels and only use L_{soft} (Kullback-Leibler divergence loss between the soft labels and the student model's output posteriors) to train the distilled models for n epochs. Subsequently, we capture snapshots of the distilled model's parameters at different epochs, resulting in a series of snapshots $s_1, s_2, ..., s_n$, which mimic the original model's training process. Recognizing the significance of membership information contained in the original model's output posteriors, we include the original model as an additional supplement in the snapshots series (denoted as s_{n+1}). While we can obtain the shadow model's training process, it does not match the exact distillation process of the target model. Distilled models converge faster with sufficient distillation data. Consequently, to align the membership information depicted in the training processes of both target and shadow models, we proceed by distilling the shadow model further, aiming to emulate similar training processes.

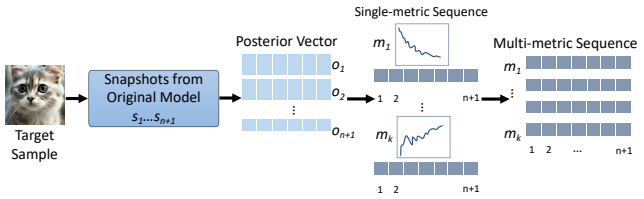


Figure 4: Workflow of multi-metric sequence construction, which assembles the membership information of a sample into a sequence of a k-dimensional space.

Construction of Multi-metric Sequences. Our construction method involves serialized feature engineering to encode membership information leaked from the output posteriors of $n+1$ snapshots into sequences in a high-dimensional space. To achieve this, we feed a given sample into $s_1, s_2, ..., s_n, s_{n+1}$ and obtain $n+1$ output posteriors $o_1, o_2, ..., o_n, o_{n+1}$. Subsequently, we calculate k metric values (e.g., Loss, Max, SD, etc., as mentioned in Section 2.2) for each output posterior o_i. These $n+1$ values of the same metric are concatenated together in temporal order, forming a single metric sequence, which becomes a $(n+1)$-dimensional vector. Finally, we concatenate the k metric sequences together as a sequential membership signal in a k-dimensional space called *multi-metric sequence*, represented as a $k * (n+1)$ matrix. See Figure 4 for an illustration of how to construct the multi-metric sequence.

Serialization Processing. As the shadow model and its distilling process are fully controlled by the adversary, they label the multi-metric sequence obtained from D^s_{train} as 1 (members), and that from D^s_{test} as 0 (non-members). Subsequently, the adversary constructs a binary dataset and uses it to train the attack model. As aforementioned, the MLP model ignores sequential or time-series information in the multi-metric sequence, which may cause the loss of some membership signals. Therefore, we utilize a recurrent neural network (RNN) attack model to process this sequential data. Specifically, to adequately emphasize the significance of different points in the multi-metric sequence, we employ an attention-based RNN as the attack model. This choice allows us to capture contextual semantics by learning weights that highlight key points in such signals for membership inference. We train the attack model by minimizing the cross-entropy loss for the binary classification task.

Membership Inference. With the trained attack model, the adversary can perform membership inference on a given target sample by following these steps: First, the target sample is encoded into multi-metric sequence by feeding it to the series of snapshots $s_1, s_2, ..., s_n, s_{n+1}$ which are from the target model. Then, this sequence can be fed into the attack model to predict its membership status, i.e., 1 or 0.

4 Experimental Setup

4.1 Datasets

We consider seven benchmark datasets of different tasks, sizes, and complexity to conduct our experiments. Concretely, we adopt four computer vision datasets, namely CIFAR10 [26], CIFAR100 [26], CINIC10 [11], GTSRB [54], and three non-computer vision datasets,

Table 2: Performance of target models, wherein training/testing accuracy is reported for each model.

Target model	CIFAR10	CIFAR100	CINIC10	GTSRB
VGG-16	1.000/0.756	1.000/0.296	1.000/0.569	1.000/0.923
ResNet-56	0.987/0.662	0.998/0.243	0.972/0.472	1.000/0.930
WideResNet-32	0.991/0.710	0.976/0.371	0.952/0.502	0.999/0.912
MobileNetV2	0.986/0.667	0.998/0.218	0.972/0.463	1.000/0.917
Target model	News	Purchase	Location	
MLPs	0.976/0.663	1.000/0.716	1.000/0.568	

namely Purchase [1], News [2] and Location [3]. See details in Appendix A.

Following [34], we divide each dataset into five parts: target training/testing dataset (D^t_{train} / D^t_{test}), shadow training/testing dataset (D^s_{train} / D^s_{test}), and distillation dataset D^k. Among them, D^s_{train}, D^s_{test} and D^k are disjoint subsets of the auxiliary dataset D^a held by the adversary. Specifically, the data partitioning is such that the sizes of the former four datasets are kept exactly the same, and the remaining data samples are placed into the distillation dataset (see details of data splitting in Appendix Table 13).

4.2 Models

For image datasets, we adopt WideResNet-32 [67], VGG-16 [52], MobileNetV2 [48] and ResNet-56 [20], as our target models. For the non-image datasets, we adopt a 2-layer MLP as the target model. These models are trained from 80 to 150 epochs due to the complexity of model architectures and datasets. For distillation, the epoch number is set to 50. The optimization algorithm used is SGD, with a learning rate ranging from 0.01 to 0.1. See the target models' performance in Table 2. Lastly, both the architecture of the shadow model and the distilled models in our experiments remain consistent with that of the target model. Note that since recent research [10] has shown that data augmentation increases membership leakage, all attack methods in this paper, including ours, are performed on target models without data augmentation.

4.3 Baselines

To demonstrate the effectiveness of `SeqMIA`, we compared it with the following MIA methods.

Shadow Training. Shadow training is a method proposed by Shokri et al. [51], which uses multiple shadow models to mimic the target model and assigns membership labels to the output posteriors from shadow models. With a large number of labeled output posteriors, it is feasible to train an attack model. Further, Salem et al. [47] employ only one shadow model to improve this method and achieve similar attack performance. In this study, the improved shadow training method [47] is adopted as one of our baselines, denoted as ST.

Metric-based Attack. Metric-based attack [53] is performed directly based on some metric values calculated from the output posteriors of the target model, and it does not require training the attack model. In this paper, we choose two metrics, prediction entropy, and modified prediction entropy, for the baseline attacks, which are denoted as MBA(Entropy) and MBA(M-Entropy), respectively.

LiRA. LiRA [6] trains N reference models, of which $N/2$ are IN models (trained with the target sample), and $N/2$ are OUT models (trained without the target sample). Then, it calculates the Gaussian distributions of losses on the target sample for IN models and OUT models. Finally, it measures the likelihood of the target sample's loss (output by the target model) under each of the distributions, and returns whichever is more likely (i.e., member or non-member). Since the online attack of LiRA requires training new IN models for every (batch of) target samples, we use its offline version in our main evaluation, denoted as LiRA. We also provide a comparison between its online version, denoted as LiRA (online), and our method on the part of datasets and models.

EnhancedMIA. EnhancedMIA [65] utilizes N distilled models to capture the loss distribution of the target sample. Since these N distilled models are trained on an auxiliary dataset relabeled with the target model, this approach eliminates the uncertainty with regard to the training set and the target sample.

TrajectoryMIA. TrajectoryMIA [34] is another state-of-the-art attack method, which exploits membership information leaked from the training process of the target model.

Among the aforementioned methods, the first two attacks represent conventional approaches that utilize the output posteriors. LiRA and EnhancedMIA are two SOTA attacks that employ multiple reference/distilled models to calibrate membership information derived from the output posteriors. Additionally, TrajectoryMIA is another SOTA attack that leverages supplementary membership signals in conjunction with the output posterior.

Lastly, when performing LiRA and EnhancedMIA, we follow previous work [6] and set $N = 64$ for image datasets and $N = 256$ otherwise.

4.4 Evaluation Metrics

First, we adopt two average-case metrics, balanced accuracy and AUC, which have been widely used in [7, 10, 19, 22, 24, 33, 43].

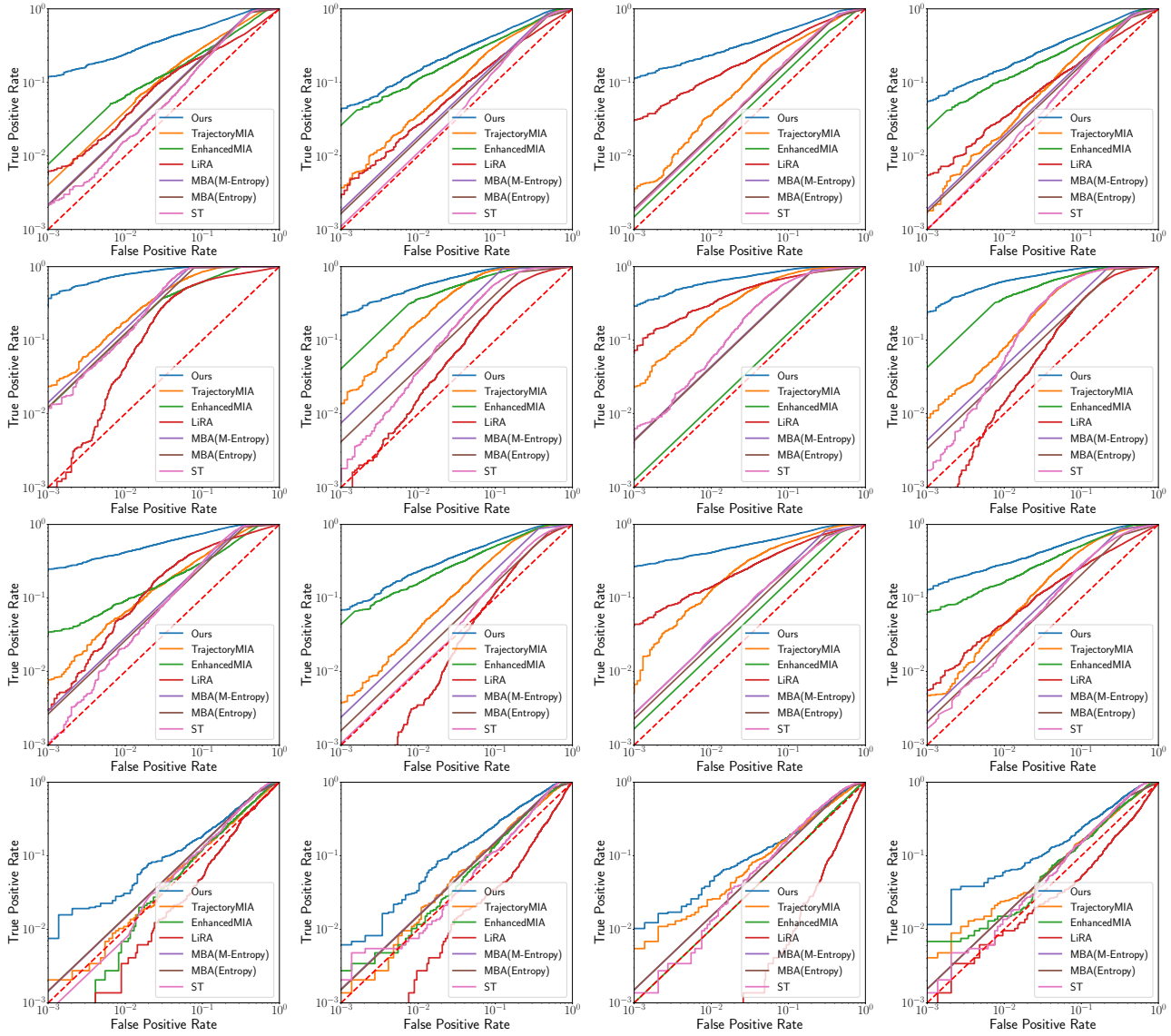
Balanced accuracy. Balanced accuracy is the probability that a membership inference attack makes a correct prediction on a balanced dataset of members and non-members.

AUC. AUC is the area under the receiver operating characteristic (ROC) curve, which is formed by the true-positive rate (TPR) and false-positive rate (FPR) of a membership inference attack for all possible thresholds.

Further, we use TPR @ low FPR and Full log-scale ROC as another two metrics recently proposed by Carlini et al. [6]. This is due to that a reliable inference attack targeting a small number of samples in the entire dataset should be taken seriously. Meanwhile, the TPR at high FPR is unreliable to the adversary. Therefore, these metrics have been used in recent works [34, 65] to evaluate the utility of MIAs more comprehensively.

TPR @ low FPR. TPR @ low FPR reports the true-positive rate at a single low false-positive rate (e.g., 0.1% FPR), which allows for a quick review of the attack performance on a small portion of samples in the entire dataset.

Full log-scale ROC. Full log-scale ROC highlights TPRs in low FPR regions by drawing the ROC curves in logarithmic scale, which provides a more complete view of attack performance than TPR @ low FPR.



(a) VGG-16 (b) ResNet-56 (c) WideResNet-32 (d) MobileNetV2

Figure 5: Log-scale ROC curves for attacks on different model architectures and four image datasets (from top to bottom: CIFAR10, CIFAR100, CINIC10, and GTSRB).

5 Experimental Results

5.1 Attack Performance

The attack performance of our SeqMIA and baseline attacks is presented in Figure 5 and Figure 6. First, we observe that SeqMIA achieves the best performance in almost all cases. Specifically, for TPR @ 0.1% FPR shown in Table 3, SeqMIA demonstrates an order of magnitude improvement compared to the baseline attacks. Regarding the two averaged metrics, balanced accuracy, and AUC, we can also find that SeqMIA outperforms all baseline attacks in most cases. Additional results can be found in our technique report [28].

Furthermore, even on the well-generalized model, SeqMIA exhibits a notable advantage over other baseline attacks in terms of TPR @ 0.1% FPR. For instance, VGG-16 trained on GTSRB achieves training and testing accuracies of 1.000 and 0.923, respectively, indicating a well-generalized target model (see Table 2). For this model, we surprisingly find that two state-of-the-art attacks, LiRA and EnhancedMIA, only achieve 0% TPR @ 0.1% FPR, as shown in Table 3. The state-of-the-art attack, TrajectoryMIA, only achieves a TPR of 0.21% at an FPR of 0.1%. In contrast, our SeqMIA demonstrates an impressive 0.75%. This superior performance can be attributed to its ability to capture and leverage the integrated membership signals:

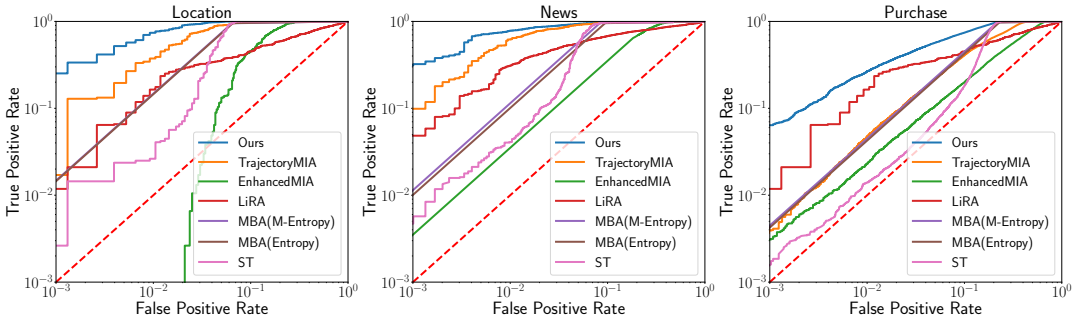


Figure 6: Log-scale ROC curves for attacks against MLPs trained on three non-image datasets.

Table 3: Attack performance of different attacks against VGG-16 trained on four image datasets.

MIA method	TPR @ 0.1% FPR (%)				Balanced accuracy				AUC			
	CIFAR10	CIFAR100	CINIC10	GTSRB	CIFAR10	CIFAR100	CINIC10	GTSRB	CIFAR10	CIFAR100	CINIC10	GTSRB
ST	0.22	1.19	0.11	0.08	0.726	0.923	0.723	0.570	0.753	0.966	0.833	0.635
MBA(Entropy)	0.21	1.23	0.27	0.15	0.735	0.945	0.789	**0.613**	0.735	0.945	0.788	0.613
MBA(M-Entropy)	0.22	1.43	0.29	0.14	0.747	0.952	0.814	0.606	0.747	0.952	0.814	0.606
LiRA	0.62	0.10	0.31	0.00	0.565	0.775	0.707	0.508	0.575	0.819	0.756	0.480
EnhancedMIA	0.77	1.21	3.45	0.00	0.636	0.836	0.717	0.567	0.694	0.904	0.767	0.575
TrajectoryMIA	0.40	2.36	0.77	0.21	0.666	0.892	0.730	0.540	0.739	0.949	0.811	0.560
SeqMIA	**11.99**	**37.03**	**24.70**	**0.75**	**0.766**	**0.959**	**0.850**	0.577	**0.869**	**0.992**	**0.937**	**0.649**

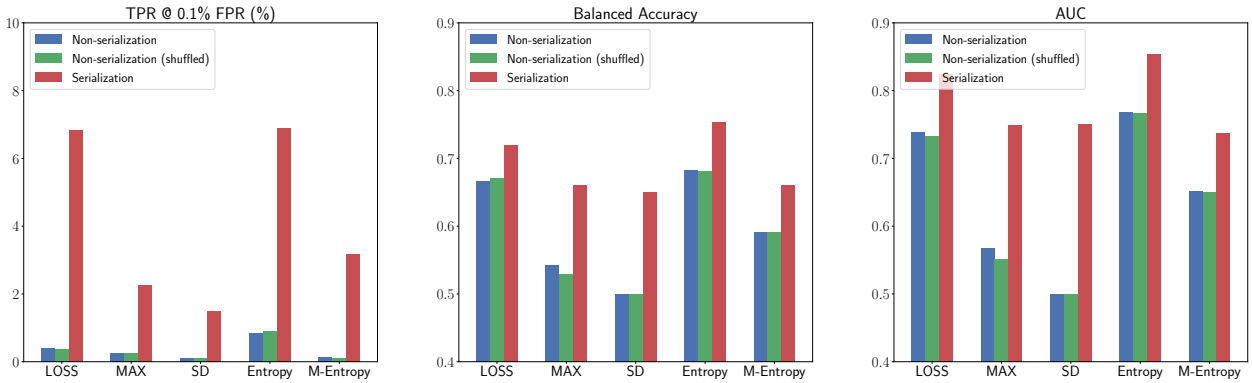


Figure 7: Performance of attacks using serialized and non-serialized membership signals against VGG-16 trained on CIFAR10.

Table 4: Attack performance of `SeqMIA` and LiRA (online) against VGG-16 trained on CIFAR10 and CIFAR100.

MIA method	TPR @ 0.1% FPR (%)		Balanced accuracy		AUC	
	CIFAR10	CIFAR100	CIFAR10	CIFAR100	CIFAR10	CIFAR100
LiRA(online)	6.38	19.91	0.687	0.884	0.766	0.961
SeqMIA	11.99	37.03	0.766	0.959	0.869	0.992

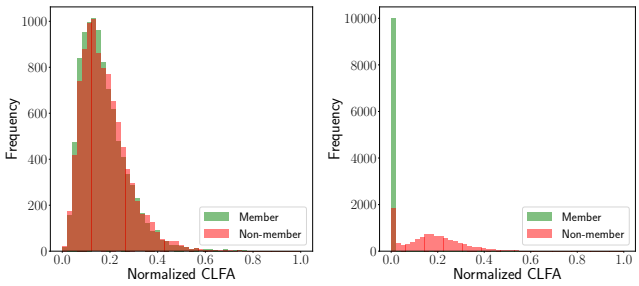
Pattern of Metric Sequence, even in scenarios where the model is well-generalized.

Lastly, as shown in Table 4, even when compared to the costly but effective method LiRA(online), `SeqMIA` consistently outperforms it, especially regarding TPR @ 0.1% FPR and balanced accuracy. This comparison further emphasizes the effectiveness of `SeqMIA`.

5.2 Analysis

Recall that we use a simple yet effective approach, the attention-based RNN, as our attack model to process the multi-metric sequences. This method automatically uncovers various patterns in the metric sequences without requiring prior characterization. While our `SeqMIA` has shown superior performance, we further investigate its success, particularly focusing on whether the attention-based RNN can indeed uncover the different patterns in the metric sequence as claimed.

Serialization vs. Non-serialization. We first investigate whether `SeqMIA` can indeed distinguish the pattern of metric sequences between members and non-members. To mitigate the effects of multiple metrics, we use only one metric at a time. In particular, we consider our `SeqMIA`, which utilizes an attention-based RNN attack



(a) First 10 Epochs (b) Last 10 Epochs

Figure 8: Distribution of CLFA for members and non-members in different training epochs for VGG-16 trained on CIFAR100.

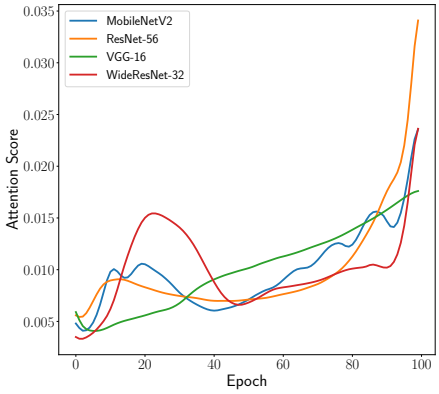


Figure 9: Attention score of our attack against four models trained on CIFAR100.

model to serialize the input, referred to as Serialization. For comparison, we also consider TrajectoryMIA, which uses an MLP attack model, referred to as Non-serialization. Additionally, we introduce a variant of TrajectoryMIA where the order of the metrics is randomized, denoted as Non-serialization (shuffled). As shown in Figure 7, we can observe that both Non-serialization and Non-serialization (shuffled) achieve similar performance in all cases. These results indicate that the MLP attack model treats the input values as independent and ignores any sequential or time-series information present in the input vector. In contrast, Serialization achieved significantly better attack performance in all cases. These results suggest that attention-based RNNs processing either sequence data or time-series data do discover patterns of metric sequences between members and non-members.

Now, we further discuss why we should adopt the attention mechanism. As shown in Figure 8, the magnitude of the loss value fluctuations for both members and non-members is large in the first 10 training epochs (i.e., the model is in an underfitting state at this stage), and thus it is difficult to distinguish between them. However, when the model is overfitted or close to being overfitted (the last 10 training epochs), members reduce the magnitude of loss fluctuations, while non-members do not. For example, almost all members

Table 5: TPR @ 0.1% FPR of RNN-based SeqMIA, Transformer-based SeqMIA, and TrajectoryMIA against ResNet-56 trained on four image datasets.

Dataset	TPR @ 0.1% FPR (%)		
	RNN	Transformer	TrajectoryMIA
CIFAR10	**4.43**	2.55	0.37
CIFAR100	**21.67**	0.65	1.38
CINIC10	**6.89**	5.94	0.38
GTSRB	**0.62**	0.62	0.14

Table 6: TPR @ 0.1% FPR for attacks using different membership signals against VGG-16 trained on four image datasets.

MIA method	TPR @ 0.1% FPR (%)			
	CIFAR10	CIFAR100	CINIC10	GTSRB
Loss set	0.40	2.36	0.77	0.21
Multi-metric set	3.92	3.09	1.12	0.21
Loss sequence	6.83	25.75	16.33	0.14
Multi-metric sequence	**11.99**	**37.03**	**24.70**	**0.75**

have loss fluctuations of less than 0.01, whereas more than half of the non-members have fluctuations of more than 0.01. We believe that this is because, at this stage, the model matches the individual characteristics of the members so that they exhibit consistently small losses. Therefore, we introduce an attention mechanism to focus on key parts of the metric sequence. Figure 9 describes the attention scores of SeqMIA with the four models trained on CIFAR100, which implies that SeqMIA is able to capture the membership signal in the metric sequences accurately.

In addition to RNNs, we further explore applying Transformer [56], a self-attention-based technique, for serializing metric sequences in SeqMIA. Transformers allow parallel processing of input sequences, offering efficiency and scalability compared to the sequential processing of RNNs. However, in our evaluation comparing RNN-based and Transformer-based attack models for SeqMIA, surprisingly, the RNN-based model performs better. We attribute this to sequence length constraints, smaller point dimensions, and the limited amount of sequences, hindering the Transformer's performance. Despite this, the Transformer-based model in SeqMIA generally surpasses TrajectoryMIA, showcasing the effectiveness of serialization in capturing membership information, as shown in Table 5. See more results in our technique report [28].

More Signals of Multiple Metric Sequences. The previous studies [12, 43] have demonstrated the potential of utilizing multiple metrics to enhance performance. However, these studies focus on non-serialized metric values in white-box scenarios and do not consider the influence of serializing multiple metrics. Here, we delve deeper into the impact of extra information from multi-metric sequence, which is initially proposed by SeqMIA.

First, we take TrajectoryMIA as the example (denoted as *Loss Set*) and extend it with multiple metrics (including loss and other metrics in Section 2.2), denoted as *Multi-metric Set*. Both approaches use a set of non-serialized metric values, which is constructed into a vector and fed into an MLP attack model for inference. Besides, we denote our SeqMIA as *Multi-metric Sequence*, and its single metric version as *Loss Sequence*. As shown in Table 6, the multi-metric set

Table 7: The impact of the number of distillation epochs for VGG-16 trained on CIFAR10.

	Distillation epochs					
	5	10	20	30	40	50
TPR @ 0.1% FPR (%)	1.99	5.75	9.83	10.60	12.37	11.99
Balanced accuracy	0.752	0.762	0.765	0.760	0.766	0.766
AUC	0.829	0.857	0.864	0.865	0.868	0.869

demonstrates higher attack performance than the loss set, which is consistent with the conclusions in [12, 43]. Interestingly, the multi-metric sequence exhibits a larger performance gain compared to the improvement achieved by the multi-metric set. For instance, when evaluating VGG-16 trained on CIFAR100, the multi-metric sequence achieves a notable 34.67% TPR @ 0.1% FPR improvement over the loss set, while the multi-metric set only improves by 0.73%. See more results in our technique report [28]. We attribute this to the fact that target models are optimized for member samples such that members will get better on multiple metrics simultaneously, whereas non-members do not.

To validate this hypothesis, we calculate the correlation matrix of multi-metric sequences (as depicted in Figure 2). We can observe that the correlation coefficients for members are usually greater than for non-members. Furthermore, we evaluate the attack performance of `SeqMIA` using dual-metric sequence, as shown in Figure 10. We can find that the best attack performance is often achieved by the two metrics, which have large differences in correlation coefficients between members and non-members. For instance, the correlation coefficient of Loss and SD for members is 0.81, whereas, for non-members, it is 0.28, as demonstrated in Figure 2. Meanwhile, Loss and SD achieves the best performance(33% TPR @ 0.1% FPR), as shown in Figure 10. See more results in our technique report [28]. Therefore, we argue that multiple metric sequences indeed contain additional membership information than single metric sequences and can further improve the attack performance.

5.3 Ablation Study

In this section, we investigate the impact of several important factors on the attack performance of our method.

Number of Distillation Epochs. The number of epochs utilized for knowledge distillation significantly impacts both the computational cost in the distillation process and the input dimension to the attack model. Therefore, it is crucial to determine the optimal number of epochs required in the distillation process.

Table 7 illustrates the impact of the number of distillation epochs on the attack performance. It is evident that increasing the number of distillation epochs can significantly increase the TPR @ 0.1% FPR, while having minimal effect on the balanced accuracy and AUC. This observation suggests that our attack is capable of distinguishing between members and non-members more reliably. As argued in [6], average metrics are often uncorrelated with low FP success rates. While the two average metrics (balanced accuracy and AUC) of our attack no longer grow significantly after 20 distillation epochs, the TPR @ 0.1% FPR continues to improve. The continued improvement of TPR @ 0.1% FPR suggests that on a small portion of samples in the entire dataset, our attack becomes more reliable. This situation should be taken into account by the model stakeholders. Additionally, the best attack performance is achieved within approximately 50 epochs, indicating that the computational cost can be effectively controlled within an acceptable range.

Table 8: The impact of distillation dataset size for VGG-16 trained on CIFAR10. The accuracy of target model is 0.569.

	Distillation dataset size					
	10k	20k	70k	120k	170k	220k
Distilled accuracy	0.564	0.566	0.563	0.561	0.574	0.566
TPR @ 0.1% FPR (%)	9.38	14.83	18.22	20.96	20.98	22.57
Balanced accuracy	0.835	0.839	0.845	0.843	0.844	0.843
AUC	0.922	0.927	0.933	0.934	0.935	0.936

Table 9: The impact of the overfitting level of the target model. The experiments are conducted on VGG-16 trained on CINIC10.

	Training dataset size				
	30k	25k	20k	15k	10k
Overfitting level	0.335	0.359	0.372	0.408	0.431
TPR @ 0.1% FPR (%)	8.41	10.21	11.85	13.03	20.43
Balanced accuracy	0.744	0.766	0.776	0.793	0.855
AUC	0.844	0.867	0.879	0.893	0.943

Size of Distillation Dataset. For knowledge distillation, the size of the distillation dataset is a crucial factor that significantly impacts the distillation performance. To investigate the influence of this factor on our attack performance, we conduct experiments with varying sizes of the distillation dataset.

We present the results in Table 8. Similarly, we observe that a larger distillation dataset size leads to higher TPR @ 0.1% FPR, while having little impact on the balanced accuracy and AUC. This finding demonstrates that a larger distillation dataset is advantageous in improving the attack performance. Besides, it further supports the claim that our attack becomes very reliable on a small portion of samples in the entire dataset as the size of distillation dataset increases.

Overfitting Level of the Target Model. It is widely acknowledged that the success of membership inference attacks is closely related to the overfitting level of the target model [47, 51]. Here we quantify the overfitting level using the training and testing accuracy gap and manipulate it by varying the size of the training set. Concretely, the distillation dataset size is kept fixed at 100,000 samples, while we manipulate the size of the target/shadow training and testing datasets, ranging from 30,000 down to 10,000 samples.

As described in Table 9, we observe that as the overfitting level increases, the attack performance improves regarding TPR @ 0.1% FPR, balancing accuracy, and AUC. Furthermore, we highlight that even when the target model exhibits good generalization with a low overfitting level (0.335), `SeqMIA` still achieves a significant 8.41% TPR @ 0.1% FPR. Surprisingly, this performance outperforms that of all baselines, even in the more overfitting scenario (overfitting level of 0.431), as demonstrated in Table 3.

Disjoint Datasets. We relax our previous assumption that the adversary possesses knowledge of the target model's training dataset

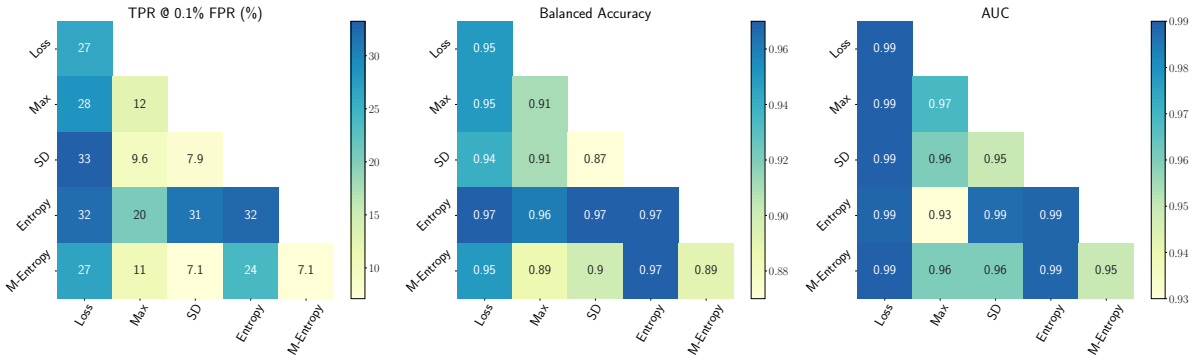


Figure 10: Attack performance of SeqMIA using double metrics against MLPs trained on Location.

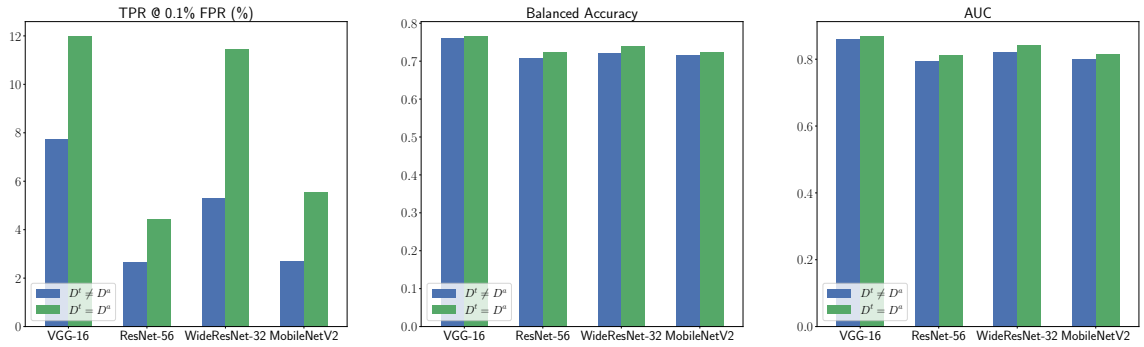


Figure 11: Attack performance of SeqMIA against different models trained on CIFAR10 by using ImageNet part ($D^t \neq D^a$) and CIFAR10 part ($D^t = D^a$) of CINIC10.

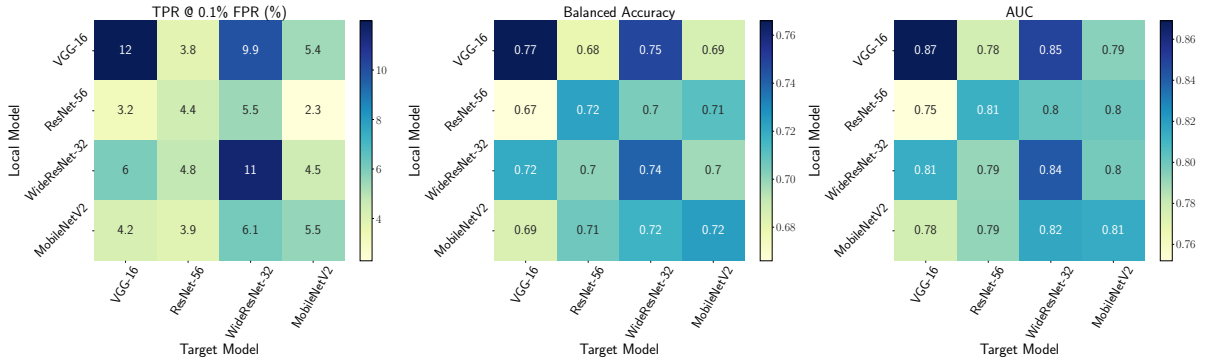


Figure 12: The impact of different model architectures used for the local (shadow and distilled) models. The target models and local models are trained on CIFAR10.

distribution. Instead, we utilize CIFAR10 as the training dataset for the target model (denoted as D^t) and a subset of CINIC10 derived from ImageNet as the dataset held by the adversary (denoted as D^a). In Figure 11, we observe that when $D^t \neq D^a$, the attack performance of SeqMIA is compromised. This degradation occurs because the discrepancy in data distribution leads to differences in prediction behavior between the target model and the models trained by the adversary, consequently affecting the quality of our constructed multi-metric sequences in capturing membership information. Nonetheless, SeqMIA's performance still surpasses that of all baselines. For instance, SeqMIA($D^t \neq D^a$) achieves a TPR @ 0.1% FPR of more than 7% against VGG-16 (as shown in Figure 11), while all baselines ($D^t = D^a$) achieve at most 0.77% (as referred to in Table 3).

Table 10: TPR @ 0.1% FPR of SeqMIA against VGG-16 trained on CIFAR10 with DP-SGD.

	δ = 1e-5 and C = 1		Accuracy of the target model	TPR @ 0.1% FPR (%)
	σ	ϵ		
No defense	-	-	0.756	11.99
DP-SGD	0	∞	0.575	0.76
	0.2	1523	0.581	0.31
	0.5	43	0.482	0.21
	1	6	0.377	0.17

Table 11: TPR @ 0.1% FPR of different attacks against VGG-16 trained on CIFAR10 with DP-SGD (σ = 0.2 and σ = 0.5).

	TPR @ 0.1% FPR (%)	
	σ = 0.2	σ = 0.5
ST	0.11	0.04
MBA(Entropy)	0.11	0.10
MBA(M-Entropy)	0.11	0.10
LiRA	0.02	0.04
EnhancedMIA	0.10	0.12
TrajectoryMIA	0.15	0.16
SeqMIA	**0.31**	**0.21**

Different Model Architectures and Hyperparameters. We proceed to relax the second assumption that requires the adversary to possess knowledge of the target model's architecture and hyperparameters. In other words, the adversary is now allowed to utilize different model architectures and hyperparameters to locally train the shadow model and distilled model. As depicted in Figure 12, the attack performance is typically optimal along the diagonal. This can be attributed to the fact that using the same model architecture and hyperparameters enables the adversary to more accurately simulate the training process of the target model. While adopting a different model architecture with different hyperparameters leads to a decrease in SeqMIA's performance, its worst-case performance still surpasses that of all baselines (when both the target and adversary models share the same architecture and hyperparameters). For instance, the worst TPR @ 0.1% FPR achieved by SeqMIA against VGG-16 is 3.2% (as shown in Figure 12), while all baselines achieve at most 0.77% (as referred to in Table 3).

6 Discussion

In this section, we evaluate the performance of SeqMIA against several existing defenses. Then, we discuss the limitations of SeqMIA.

6.1 Defense Evaluation

To mitigate the risk of membership leakage, a large body of defense mechanisms have been proposed in the literature [4, 25, 29, 42, 44, 62, 70]. In this section, we thoroughly evaluate the effectiveness of SeqMIA against three prominent defenses, namely DP-SGD [4], Adversarial Regularization [42], and MixupMMD [29].

DP-SGD. Table 10 presents the performance of SeqMIA under DP-SGD, evaluated on VGG-16 trained on CIFAR10 (see more results in our technique report [28]). We employ the Opacus library to implement DP-SGD and fix the parameters δ = 1e-5 and C = 1, following [6, 34]. We can observe that the attack performance of SeqMIA gradually decreases as the defense effects increase. However, stronger defense effects also result in a sharp drop in the accuracy of the target model. To balance defense strength and model accuracy, we select the cases of σ = 0.2 and σ = 0.5 for further analysis. These settings offer acceptable trade-offs between defense strength and model accuracy. Table 11 shows that SeqMIA still outperforms other baselines under DP-SGD. For instance, when σ = 0.2, the TPR @ 0.1% FPR of SeqMIA is more than twice that of other baselines.

Table 12: TPR @ 0.1% FPR of different attacks against ResNet-56 trained on CINIC10 with AdvReg and MixupMMD.

	TPR @ 0.1% FPR (%)		
	no defense	AdvReg	MixupMMD
ST	0.11	1.08	0.06
MBA(Entropy)	0.16	0.12	0.25
MBA(M-Entropy)	0.24	0.52	0.23
LiRA	0.00	0.02	0.11
EnhancedMIA	4.39	2.27	0.38
TrajectoryMIA	0.38	3.08	1.81
SeqMIA	**6.89**	**24.62**	**4.62**

Adversarial Regularization. Adversarial Regularization (AdvReg) is an adversarial training-based defense that adds noise to the output posteriors, making it challenging for adversaries to distinguish between members and non-members. As demonstrated in Table 12 (see more results in our technique report [28]), SeqMIA continues to achieve the best attack performance in almost all cases. Interestingly, we observe that AdvReg's co-training with the target model results in members being more involved in the training of the target model, which makes them more significantly different from non-members. Thus, both SeqMIA and TrajectoryMIA demonstrate enhanced attack performance. Notably, this enhancement is more pronounced for SeqMIA, as it leverages more membership signals leaked from the training process. For instance, when the target model has no defense, SeqMIA achieves a TPR @ 0.1% FPR of 6.89%, and when the target model is protected by AdvReg, the TPR @ 0.1% FPR increases to 24.62%.

MixupMMD. MixupMMD is a defense aimed at mitigating membership inference attacks by reducing the target model's generalization gap. As previously discussed, the overfitting level of the target model plays a crucial role in membership leakage. Consequently, MixupMMD leads to a degradation in the performance of all attacks, including our SeqMIA, as depicted in Table 12 (see more results in our technique report [28]). However, it is worth noting that despite this degradation, SeqMIA continues to outperform other baseline attacks in almost all cases.

6.2 Limitations

SeqMIA has limitations as follows: it cannot be applied to label-only scenarios due to its reliance on the output posterior, and it is not suitable for large model scenarios regarding computation because it requires training and distilling the shadow model. Therefore, model holders can only provide predicted labels instead of the posterior to defend against SeqMIA.

7 Related Works

7.1 Membership Inference Attacks

Nowadays, there exist a wide range of other security and privacy research in the machine learning domain [8, 17, 18, 21, 30, 32, 35, 36, 39, 40, 46, 49, 50, 58–60, 64, 68]. In this work, we mainly focus on membership inference attacks. Membership inference attacks have been successfully performed in various settings about the adversary's knowledge, including white-box [27, 43], black-box [9, 24, 47, 51, 53, 69], and label-only [10, 33] settings. They have been applied in many machine learning scenarios, such as federated learning [41, 43, 55] and multi-exit networks [31], etc.

Specifically, Shokri et al. [51] and Salem et al. [47] proposed a shadow training technique that employs shadow models to acquire the membership signals. Moreover, Song et al. [53] and Yeom et al. [66] proposed the metric-based attack that directly compares losses or other metric values of samples with a predefined threshold. In addition, some membership signals obtained in the white-box scenario are incorporated to improve the attack performance [12, 43]. Besides, label-only attacks [10, 33, 61] solely rely on the predicted labels to acquire the membership signals. Recently, researchers [6, 34, 45, 57, 65] focused on reducing the false positives of MIAs by using each sample's hardness threshold to calibrate the loss from the target model. Further, Bertran et al. [5] proposed a new attack via quantile regression, which can obtain performance close to that of LiRA [6] with less computation. Moreover, Liu et al. [34] presented TrajectoryMIA, which utilizes the membership signals generated during the training of the target model.

7.2 Defenses Against MIAs

Since the overfitting level is an important factor affecting membership leakage, some regularization techniques have been used by [37, 47, 51] to defend against membership inference attacks, such as L2 regularization, dropout and label smoothing, etc. Recently, Li et al. [29] proposed the method MixupMMD to mitigate membership inference attacks by reducing the target model's generalization gap. Furthermore, Abadi et al. [4] proposed a more general privacy-preserving method DP-SGD, which adds differential privacy [14] for the stochastic gradient descent algorithm. Subsequently, some works [44, 62, 70] focus on reducing the privacy cost of DP-SGD through adaptive clipping or adaptive learning rate. In addition, for membership inference attacks, some elaborate defense mechanisms, such as AdvReg [42] and MemGuard [25], have been conceived to obscure the differences between the output posteriors of members and non-members.

8 Conclusion

In this work, we introduce a new, more integrated membership signal: *the Pattern of Metric Sequence*, which comes from various stages of model training. We verify that this new signal not only includes these existing well-applied signals but also pays more attention to time-dependent patterns, such as fluctuations and correlations. Based on this new signal, we propose a novel membership inference attack against ML models, named Sequential-metric based Membership Inference Attack. We construct sequential versions of multiple metrics obtained from the training process of the target model (multi-metric sequences) and leverage an attention-based RNN to automatically mine the patterns of the metric sequences for inference. Extensive experiments demonstrate `SeqMIA` outperforms advanced baselines. we further conduct in-depth comparative analyses of metric non-sequences vs. metric sequences, and single vs. multiple metrics, revealing the reasons for its superior performance. Then, we analyze some other factors on the attack performance. Additionally, we demonstrate that `SeqMIA` outperforms existing advanced baseline attacks under several representative defenses. In the future, we aim to explore enhanced metrics with richer membership information and employ more efficient serialization models to further improve membership inference performance.

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).

References

[1] https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data.
[2] http://people.csail.mit.edu/jrennie/20Newsgroups.
[3] https://sites.google.com/site/yangdingqi/home/foursquare-dataset.
[4] 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.
[5] 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.
[6] 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.
[7] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 343–362. Association for Computing Machinery, 2020.
[8] Joann Qiongna Chen, Xinlei He, Zheng Li, Yang Zhang, and Zhou Li. A comprehensive study of privacy risks in curriculum learning. *arXiv preprint arXiv:2310.10124*, 2023.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[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] 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.
[15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333. Association for Computing Machinery, 2015.
[16] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 619–633. Association for Computing Machinery, 2018.

[17] Ge Han, Zheng Li, Peng Tang, Chengyu Hu, and Shanqing Guo. Fuzzgan: A generation-based fuzzing framework for testing deep neural networks. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 1601–1608. IEEE, 2022.

[18] Ge Han, Ahmed Salem, Zheng Li, Shanqing Guo, Michael Backes, and Yang Zhang. Detection and attribution of models trained on generated data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4875–4879. IEEE, 2024.

[19] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

[20] 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.

[21] Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. Membership-doctor: Comprehensive assessment of membership inference against machine learning models. *arXiv preprint arXiv:2208.10445*, 2022.

[22] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Reconstruction and membership inference attacks against generative models. *arXiv preprint arXiv:1906.03006*, 2019.

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[24] 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.

[25] 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.

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

[27] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated White-Box membership inference. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1605–1622. USENIX Association, 2020.

[28] Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. Seqmia: Sequential-metric based membership inference attack. *arXiv preprint arXiv:2407.15098*, 2024.

[29] 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.

[30] Zheng Li, Chengyu Hu, Yang Zhang, and Shanqing Guo. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of dnn. In *Proceedings of the 35th annual computer security applications conference*, pages 126–137, 2019.

[31] Zheng Li, Yiyong Liu, Xinlei He, Ning Yu, Michael Backes, and Yang Zhang. Auditing membership leakages of multi-exit networks. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 1917–1931. Association for Computing Machinery, 2022.

[32] Zheng Li, Ning Yu, Ahmed Salem, Michael Backes, Mario Fritz, and Yang Zhang. {UnGANable}: Defending against {GAN-based} face manipulation. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 7213–7230, 2023.

[33] 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.

[34] 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.

[35] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Backdoor attacks against dataset distillation. *arXiv preprint arXiv:2301.01197*, 2023.

[36] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502*, 2023.

[37] 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.

[38] 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.

[39] Yihan Ma, Zhengyu Zhao, Xinlei He, Zheng Li, Michael Backes, and Yang Zhang. Generative watermarking against unauthorized subject-driven image synthesis. *arXiv preprint arXiv:2306.07754*, 2023.

[40] Kai Mei, Zheng Li, Zhenting Wang, Yang Zhang, and Shiqing Ma. Notable: Transferable backdoor attacks against prompt-based nlp models. *arXiv preprint arXiv:2305.17826*, 2023.

[41] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[42] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 634–646. Association for Computing Machinery, 2018.

[43] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[44] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

[45] 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.

[46] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 703–718. IEEE, 2022.

[47] 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.

[48] 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.

[49] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023.

[50] Xinyue Shen, Xinlei He, Zheng Li, Yun Shen, Michael Backes, and Yang Zhang. Backdoor attacks in the supply chain of masked image modeling. 2022.

[51] 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.

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

[53] 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.

[54] 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.

[55] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 14(6):2073–2089, 2019.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[57] 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.

[58] Rui Wen, Zhengyu Zhao, Zhuoran Liu, Michael Backes, Tianhao Wang, and Yang Zhang. Is adversarial training really a silver bullet for mitigating data poisoning? 2023.

[59] Yixin Wu, Rui Wen, Michael Backes, Pascal Berrang, Mathias Humbert, Yun Shen, and Yang Zhang. Quantifying privacy risks of prompts in visual prompt learning. 2024.

[60] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. 2022.

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

[62] Zhiying Xu, Shuyu Shi, Alex X Liu, Jun Zhao, and Lin Chen. An adaptive and fast convergent approach to differentially private deep learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1867–1876. IEEE, 2020.

[63] Dingqi Yang, Daqing Zhang, and Bingqing Qu. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Trans. Intell. Syst. Technol.*, 7(3):1–23, 2016.

[64] Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pages 39299–39313. PMLR, 2023.

[65] 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.
[66] 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.
[67] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
[68] Boyang Zhang, Zheng Li, Ziqing Yang, Xinlei He, Michael Backes, Mario Fritz, and Yang Zhang. Securitynet: Assessing machine learning vulnerabilities on public models. *arXiv preprint arXiv:2310.12665*, 2023.
[69] 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.
[70] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*, 2018.

A Dataset Description

CIFAR10/CIFAR100. CIFAR10 and CIFAR100 [26] are commonly used datasets for evaluating image recognition algorithms, each including 60,000 color images of size 32×32. The difference is only that the images in CIFAR10 are equally distributed into 10 classes, while the images in CIFAR100 are equally distributed into 100 classes.

CINIC10. CINIC10 [11] contains 270,000 images within the same classes as CIFAR10. In particular, 60,000 samples belong to CIFAR10, while the other samples come from ImageNet [13].

GTSRB. GTSRB [54] is a benchmark dataset used for traffic sign recognition, which includes 51839 images in 43 classes. Since the size of these images is not uniform, we resize them to 32×32 pixels during data preprocessing.

Purchase. Purchase is a dataset of shopping records with 197324 samples of 600 dimensions, which is extracted from Kaggle's "acquire valued shopper" challenge. Following previous works [34, 47, 51], we cluster these data into 100 classes for evaluating membership inference attacks against non-image classifiers.

News. News is a popular benchmark dataset for text classification. This dataset includes 20,000 newsgroup documents of 20 classes. Following [47], we convert each document into a vector of 134410 dimensions using TF-IDF.

Location. Location is a preprocessed check-in dataset provided by Shokri et al. [51], which is obtained from Foursquare dataset [63]. Location contains 5010 data samples of 446 dimensions across 30 classes.

Table 13: Data splits for our evaluation.

Dataset	$\mathcal{D}^t_{train}$	$\mathcal{D}^t_{test}$	$\mathcal{D}^s_{train}$	$\mathcal{D}^s_{test}$	$\mathcal{D}^k$
CIFAR10	10000	10000	10000	10000	20000
CIFAR100	10000	10000	10000	10000	20000
CINIC10	10000	10000	10000	10000	220000
GTSRB	1500	1500	1500	1500	45839
Purchase	20000	20000	20000	20000	110000
News	3000	3000	3000	3000	6000
Location	800	800	800	800	1400

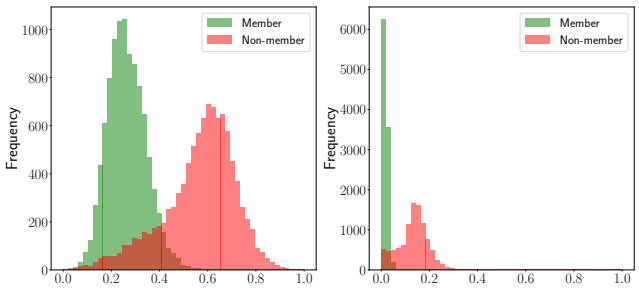


Figure 13: The distributions of the cumulative fluctuation amplitude of SD (left) and M-Entropy (right) within 100 epochs, which is obtained from 10,000 members and 10,000 non-members of VGG-16 trained on CIFAR100.

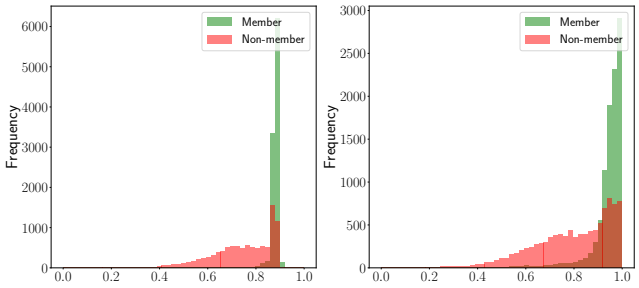


Figure 14: The distributions of the decline rate of Loss (left) and Entropy (right) within 100 epochs, which is obtained from 10,000 members and 10,000 non-members of VGG-16 trained on CIFAR100.

B Additional Time-Dependent Patterns of Metric Sequences

Decline Rate of Metric Sequences. We choose two metrics, Loss and Entropy, as our example. First, we construct the sequences of loss values and entropy values for each sample as the training progresses. And the loss decline rate for each sample is calculated by measuring the loss decline amplitude within a period of *consecutive epochs* and then dividing by the number of epochs. Similarly, we can obtain decline rate of entropy sequence. We then count the frequency of the samples regarding the distribution of their decline rate of loss (or entropy). As depicted by Figure 14, we observe members exhibit significantly larger decline rate of loss (or entropy) sequence compared to non-members. The results reconfirm that there exists a very clear difference in the pattern of metric sequence (e.g., loss sequence and entropy sequence) between members and non-members.