SoK: Membership Inference is Harder Than Previously Thought

Antreas Dionysiou
University of Cyprus
Nicosia, Cyprus
dionysiou.antreas@ucy.ac.cy

Elias Athanasopoulos
University of Cyprus
Nicosia, Cyprus
athanasopoulos.elias@ucy.ac.cy

ABSTRACT

Membership Inference Attacks (MIAs) can be conducted based on specific settings/assumptions and experience different limitations. In this paper, first, we provide a systematization of knowledge for all representative MIAs found in the literature. Second, we empirically evaluate and compare the MIA success rates achieved on Machine Learning (ML) models trained with some of the most common generalization techniques. Third, we examine the contribution of potential data leaks to successful MIAs. Fourth, we examine if the depth of Artificial Neural Networks (ANNs) affects MIA success rate and to what extent. For the experimental analysis, we focus solely on well-generalizable target models (various architectures trained on multiple datasets), having only black-box access to them.

Our results suggest the following: (a) MIAs on well-generalizable targets suffer from significant limitations which undermine their practicality, (b) common generalization techniques result in ML models which are comparably robust against MIAs, (c) data leaks, although effective for overfitted models, do not facilitate MIAs in case of well-generalizable targets, (d) deep ANN architectures are not more vulnerable to MIAs compared to shallower ones or the opposite, and (e) well-generalizable models can be robust against MIAs even when not achieving state-of-the-art performance.

KEYWORDS

Membership inference attack, adversarial machine learning, privacy

1 INTRODUCTION

Membership Inference Attacks (MIAs), where an adversary given a trained Machine Learning (ML) model and a target data record determines whether this data record was used as part of the model's training set [64], have severe consequences affecting directly users' privacy [44]. Learning that a record was used to train a particular ML model is an indication of *information leakage* through that model. Take for example an ML model trained to classify a patient's information to a specific disease class. If an adversary knows that a patient's data was included in the model's training set they can conclude about the patient's health status [60]. Similarly, inferring that statistics collected over a sensitive timeframe or sensitive locations include a particular user harms the individual's privacy [55]. Finally, MIAs can damage the ML model provider's intellectual property of the training dataset since collecting and labeling data instances may require lots of expensive (often human) resources [28].

From a different perspective, membership inference can be useful for: (a) supporting the suspicion that a model was trained on personal data without an adequate legal basis, or for a purpose not compatible with the data collection, (b) enforcing individual rights, such as the "right to be forgotten", and (c) detecting violations of data-protection regulations, such as the GDPR [16, 47]. For example, Song et al. [68] refer to their MIA as an *auditing technique* that helps users check if their data was used to train a natural-language text generation model without them knowing. In a similar fashion, Carlini et al. [6] propose a MIA-based methodology that can benefit privacy by allowing ML practitioners to quantitatively assess the risk that rare or unique training-data sequences are unintentionally memorized by generative models. Thus, membership inference can be useful not only to malicious entities, but also to legitimate entities who wish to examine if specific records have been included in an ML model's dataset without the corresponding permission.

Systematizing Prior Works. MIAs can be conducted based on specific settings/assumptions and experience different limitations (see Table 4, Appx. A.1, for an overview). Surprisingly, the set of settings/assumptions is quite diverse. For example, MIAs are well known to work on overfitted models [25, 43, 60, 79], failing, however, when the target model is *well-generalizable*. Such models classify previously unseen (testing) input samples with high success rates. In addition, MIAs achieving high success rates on well-generalizable targets might impose significant limitations, which undermine their practicality. In general, a systematic classification of MIAs found in the literature, as well as an in-depth analysis of the assumptions/limitations affecting either their generality/applicability or their performance, is currently lacking.

Evaluating ML Generalization Techniques. Training an ML model involves the use of certain optimization and regularization techniques for increasing the model's generalization. However, the extent to which different ML generalization techniques affect the MIA success rates currently remains unclear. In other words, specific generalization techniques may result in ML models that are more vulnerable to MIAs compared to other models trained using different generalization techniques.

For example, the use of a particular optimizer may result in a local optima (set of weights' values) that makes the model more vulnerable to MIAs than other models trained using different optimizers. Furthermore, the use of specific regularization techniques may increase, rather than decrease, the vulnerability to MIAs [31, 72]. This is because, while regularization may indeed help, it does not guarantee no-overfitting. Thus, applying specific regularizers may result in models that are well-generalizable yet leave particular groups of instances exposed to MIAs (e.g., due to increased distributional overfitting [35], where some records are more vulnerable to MIAs than others). In general, a detailed comparison between the MIA success rates achieved on ML models trained with different optimizers and regularizers, is currently absent from the literature.

Proceedings on Privacy Enhancing Technologies 2023(3), 286–306

https://doi.org/10.56553/popets-2023-0082

Data Leakage Implications. The adversary's knowledge of the target model's training set is another important factor affecting MIA success rates, at least on overfitted models. For example, Hui et al. [25] showed that MIAs' performance on overfitted models is analogous to the number of *labeled instances* available to the adversary. Labeled instances include the ground-truth membership status, that is, whether or not each instance was included in the target model's training set. However, it is currently unknown whether Hui et al.'s observation holds for well-generalizable targets as well. Assuming that an adversary has access to the ground-truth membership information of the target model's training set is a realistic scenario as real-world incidents, e.g., data breaches, have actually occurred several times in the past [15, 19, 30] [1]. Exploring the consequences of potential data leaks and examining their contribution to black-box MIAs (a best-case scenario for black-box adversaries with partial access to ground-truth membership status) is a case that, to the best of our knowledge, has been largely overlooked.

Depth Factor. Deep Artificial Neural Networks (ANNs) demonstrate excellent performance on a wide-range of real-life problems. Generally speaking, the more layers an ANN has, the higher its accuracy [3]. Nonetheless, deep ANNs tend to memorize the training data because they have the capacity to do so [50, 71, 82]. Thus, although not overfitted, deep ANNs *may* be more vulnerable to MIAs compared to shallow ones –still remains unclear. In general, an experimental investigation regarding whether or not, and to what extent, the depth of ANNs facilitates the successful deployment of black-box MIAs is currently absent from the literature.

Contributions. Our analysis is not only literature-based, but mostly experimentally based. In fact, we carry out several MIAs in order to explore the key adversarial settings that contribute the most in the effectiveness of the attacks. Our literature-based and experimental exploration provides several new results. The key results are the contributions of this paper, which we summarize here.

- We provide a systematization of knowledge for all representative MIAs found in the literature. Our analysis shows that the majority of MIAs report high attack success rates only on overfitted models; papers achieving high MIA success rates on well-generalizable models either rely on adversary's knowledge about the target model's internals and/or its data distribution, or impose significant limitations, which undermine their practicality (Sec. 3).
- We evaluate and compare the MIA success rates achieved on ML models trained with some of the most common ML generalization techniques. Our results suggest that all tested optimizers and regularizers lead to ML models which are *comparably* robust against black-box adversaries with strong background knowledge. In fact, we show that one can train well-generalizable models, which are robust against MIAs, without using any regularization technique. Thus, we conclude that MIA success rate is more closely related to the target model's generalization rather than to a *particular* regularization technique used. Furthermore, we observe that MIA success rate is not directly related to the model's performance rather than its generalization gap [2]. In other words, a well-generalizable model can be robust against MIAs even if *not* achieving state-of-the-art performance (Sec. 4).
- We examine if Hui et al.'s observation (i.e., MIA success rate is analogous to the number of labeled instances available to the adversary) holds for well-generalizable targets as well, apart from overfitted ones. We do this by simulating the case where a large portion of the target dataset has been leaked by hackers or insiders, and by conducting black-box MIAs against some of the most popular deep ANNs. Our results suggest that Hui et al.'s observation does *not* apply to well-generalizable targets (Sec. 5).
- Deep ANNs, although not overfitted, might memorize the training data, and thus be more vulnerable to MIAs compared to shallower ones. We, for the first time, experimentally investigate whether or not the depth of ANNs facilitates black-box MIAs and to what extent. Our results suggest that deep ANN architectures are *not* more vulnerable to black-box MIAs compared to shallower ones or the opposite (Sec. 5).
- To foster further research on this topic and ease reproducibility, we release the code for our analysis [3].

2 PRELIMINARIES

2.1 The Membership Inference Problem

MIAs infer whether a particular data record has been included in a target model's training set, D, or not. Formally, given a trained ML model, M_{tar}, a target data record, x_{tar}, and any available external knowledge of an adversary, K_{adv}, a MIA model, A, can be defined as $A : x_{tar}, M_{tar}, K_{adv} \rightarrow \{0, 1\}$, where 0 means that x_{tar} is *member* in D and 1 means the opposite. Depending on their goal, MIAs can be further divided into two groups: sample-level and user-level. The goal of user-level MIAs is to determine whether a user's data was used to train an ML model, whereas the goal of sample-level MIAs is to determine the membership status of individual data records.

In the context of MIAs value can be characterized by the attack success rate as an evaluation of what level of leakage is present in M_{tar} or what amount of knowledge an attacker can expect to gain [74]. The success rate of an attack, however, can be described using various metrics. For example, we can use *precision*, which indicates the fraction of instances inferred as members and are actually members in D –True Positives (TP), divided by the sum of TP and falsely predicted members –False Positives (FP). Considering only precision for concluding about A's performance is *not* enough since the model can gain 100% success rate by just making a handful of positive predictions (i.e., members in D) for which it is highly confident in regards to those records membership status.

Another approach is to consider *recall*, which indicates the fraction of instances inferred as members and are actually members in D (TP), divided by the total number of instances in D. Considering solely recall, however, for concluding about A's performance is *not* enough since the model can gain 100% success rate by always predicting a record as a member in D.

Reporting both precision and recall, and maybe their harmonic mean –F1 score, for each class would be necessary in case of: (a) imbalanced datasets, where the samples' distribution across the classes is not uniform [58], and (b) different costs of FP and False

[1] Note that a detailed analysis on how an adversary can gain access to such information is out of scope for this paper.

[2] That is, the gap between the training and testing accuracies [5, 67].

[3] https://bitbucket.org/srecgrp/sok-membership-inference-public/

Negatives (FN). Nonetheless, in all of our experiments, the utilized datasets are balanced, having 50% members and 50% non-members, and the cost of FP equals to the cost of FN. That is, the cost of considering a record as a member when in fact it is not, is the same as considering a record as non-member when in fact it is. Thus, similarly to [16, 49, 50, 59, 60, 64, 74], we can solely rely on *accuracy*, which indicates the fraction of predictions our model got right, for reporting the attacker's performance on conducting MIAs.

2.2 Overfitted vs. Well-Generalizable Targets

Overfitted models do not generalize well on unseen, during the training phase, data records. They essentially memorize the training records and demonstrate low performance on testing records.

On the one hand, facing overfitted models is a possible scenario since: (a) some real-world datasets are very tiny and not at all representative of the overall distribution, (b) regularization techniques may indeed help, but they do not guarantee no-overfitting, and (c) distributional overfitting may exist, where some records are more vulnerable to MIAs than others.

On the other hand, one can argue that assuming overfitted targets is simply a malpractice, and ask why would this be done in the first place. This is an intuitive question since overfitted models have *no* practical use, because they yield low performance on unseen (testing) data, and the results on such models should not be generalized to well-generalizable models [58]. Furthermore, in many cases, privacy-sensitive ML models are carefully trained to be well-generalizable since this is a requirement. Finally, the different Machine Learning as a Service (MLaaS) platforms operate increasingly well in the sense that they either offer high-performing pre-trained models for various tasks or manage to achieve high accuracy results on custom datasets uploaded by the users.

For the aforementioned reasons, in this paper, we either train/test each target model for ensuring that it is not overfitted on the training data and reached an adequate level of generalization or use popular pre-trained architectures that achieve excellent performance. As Carlini et al. [5] suggest, *this makes our analysis much more realistic than prior work, which often uses models with higher error rates than our models*. Besides, as Long et al. [44] explain, *in an overfitted model, almost all of its records are vulnerable to MIAs.*

Note that while different models might experience different levels of generalization, we are not considering this in our analysis. Specifically, we are only interested in increasing the generalization of each model as much as we can. This is because Salem et al. [60] have already demonstrated that MIA success rate is analogous to the *overfitting level* of the target model. In other words, the larger the generalization gap (overfitting), the higher the MIA performance. In addition, Yeom et al. [79] provide both experimental and theoretical evidence showing that ML models become more vulnerable to MIAs as they overfit more.

2.3 White-box vs. Black-box Adversaries

White-box adversaries have access to the internals of the target model, M_{tar}, such as its architecture, weights, hidden layers' activations, and loss function.

Black-box adversaries can only query M_{tar} with an arbitrary input, x, and receive the confidence score for each class as a response, without knowing any additional information. In other words, contrary to white-box adversaries, the parameters of M_{tar} as well as the intermediate steps of the computation are not accessible.

Fully black-box adversaries are further limited in capabilities, in the sense that they can only access M_{tar}'s predicted (discrete) class without, however, the confidence score for each class.

Although in Sec. 3 we provide a systematization of knowledge for MIAs operating in any setting, later, in Sec. 4 & 5, we conduct our experimental analysis using black-box adversaries only. The motivation behind this decision is three-fold. First, it was shown that white-box adversaries have limited advantages compared to black-box adversaries [59, 63, 69]. Second, we want to examine the possibility of conducting practical MIAs under minimal adversarial assumptions. Third, this is the setting that most MLaaS platforms operate in [50, 64]. Thus, for performing our analysis, we utilize the target model's confidence scores when querying it with member or non-member data records. More specifically, we utilize the confidence scores derived from the last (output) layer as Nasr et al. [50] revealed that it leaks the most membership information.

2.4 Dataset Complexity

The dataset's complexity largely affects the performance of MIAs. For example, tasks with higher-dimensional outputs (many classes) are more vulnerable to MIAs than those with lower-dimensional outputs [56, 60, 62, 74]. Moreover, the distribution of the training data, and the uniformity within each class, significantly impact adversaries' ability to conduct MIAs [64, 74, 79].

In this paper, we mainly utilize image datasets, namely CIFAR-10/100 [34] and SVHN [51], for conducting our analysis. The intuition behind this decision is two-fold. First, image datasets have *higher* complexity compared to non-image ones [69]. Second, we avoid using classic record-based datasets, such as Purchase, Locations, and Texas (see [64] for details), since they are *not* meaningful benchmarks for privacy because of their simplicity [5]. Instead, we select image datasets that contain rich information since doing so represents a much more challenging task for the attacker compared to simple datasets, such as MNIST, where the samples from each class have very similar features [16]. Nonetheless, for supporting the generalizability of our results, we present experiments on three additional datasets that are not image related. In particular, we utilize two tabular datasets, namely Adult [11] and Surgical [61], and one text dataset, namely IMDB [46]. Note, however, that a detailed analysis for the effects of the dataset's complexity on the MIA success rates achieved is beyond the scope of this paper.

For the image datasets (CIFAR-10/100 and SVHN), we use Convolutional Neural Network (CNN) based architectures to build the target models, whereas for the other datasets (Adult, Surgical and IMDB), we use Multilayer Perceptron (MLP).

3 SYSTEMATIZATION OF KNOWLEDGE

In this section, we provide an in-depth systematization of the related literature, highlighting the papers that achieve well-above baseline MIA success rates on well-generalizable targets, as well as their assumptions and limitations (Sec. 3.1). This helps us in identifying specific research questions that still remain unanswered (Sec. 3.2). We then answer those research questions in Sec. 4 & 5.

3.1 Analysis of Representative MIAs

MIAs operate under specific settings and assumptions, and thus experience different limitations (see Table 4, Appx. A.1, for an overview of the most common assumptions). Table 1 shows a detailed classification of the characteristics of different types of adversaries found in the literature. As shown, numerous MIAs have been proposed, each of them targeting different ML models (either discriminative or generative) or aggregate statistics in both standalone and collaborative environments. In this paper, we focus on MIAs targeting ML models due to their increased popularity and performance on solving various tasks.

As shown in Table 1, the majority of MIAs report high success rates only on overfitted models; papers that achieve moderate or high MIA success rates on well-generalizable models (see highlighted rows) either rely on the adversary's knowledge about the target model's internals, such as its structure/parameters, training algorithm, data distribution, or impose significant limitations.

Below, we first outline the most representative MIAs on overfitted targets. Later, we analyse MIAs on well-generalizable targets. For the reasons mentioned in Sec. 2.2, we give greater focus on MIAs targeting well-generalizable models. For each work, we provide a summary of its operation and then outline the downsides/limitations that affect either its practicality or effectiveness.

3.1.1 MIAs on Overfitted Targets. Shokri et al. [64] were the first to propose MIAs on (trained) discriminative ML models having only black-box access to them. Shokri et al. achieve 93.5% inference accuracy when the target model is overfitted on the training data having testing accuracy of 65%. However, their inference accuracy drops to 51.7% (baseline is 50%) when the target model is well-generalizable having testing accuracy over 90%.

Later, Salem et al. [60] further analyse the connections between overfitting and MIA success rates. However, the authors observe the same results with Shokri et al. In particular, when they target a well-generalizable model trained on Adult dataset, they achieve relatively weak performance. On the other hand, when they attack an overfitted model that demonstrates significant gap between the training and testing accuracies (more than 78%), the MIA success rate rises to 95%. Furthermore, Salem et al.'s black-box MIAs relax two rather strong assumptions made by Shokri et al. First, the utilized shadow models, which mimic the target model's behaviour, do *not* need to share the same structure as the target model. Second, the dataset used to train those shadow models does *not* need to come from the same distribution as the target model's training set.

Recently, Hui et al. [25] propose MIAs that probe the target model and extract membership semantics using a novel approach, called differential comparison. Hui et al.'s MIAs do *not* require the use of shadow models and instead infer membership directly from the probing results obtained from the target model. Nonetheless, similar to Shokri et al. [64] and Salem et al. [60], Hui et al. mainly focus on targeting overfitted rather than well-generalizable models.

3.1.2 MIAs on Well-generalizable Targets. Yeom et al. [79] conduct a formal and empirical analysis of the impact that overfitting has on an attacker's ability to carry out MIAs. The authors express the advantage of an attacker as a function of the extent of overfitting, thereby showing that a model's generalization is a strong predictor for its vulnerability to MIAs. However, Yeom et al. show that overfitting, although sufficient, is not a necessary condition for enabling MIAs, and thus motivate the study of other factors that may affect such attacks. Both Yeom et al. and Sablayrolles et al. [59] conduct successful MIAs on well-generalizable targets by exploiting the model's *loss function*. In fact, Sablayrolles et al. show (theoretically) that the optimal MIA depends solely on the loss function. Nonetheless, for specific privacy-sensitive ML models, such as those deployed in the healthcare field, the loss function may *not* be accessible to the end users. Thus, in the (common) case where an adversary can only query the target model and receive, as a response, the confidence score for each class, both Yeom et al.'s and Sablayrolles et al.'s MIAs *cannot* be applied.

Nasr et al. [50] propose MIAs in both standalone and collaborative environments. The authors show that well-generalizable models, despite achieving state-of-the-art performance on CIFAR-100, are still susceptible to *white-box* MIAs. In collaborative setting, they consider both passive and active participants showing that repeatedly updating the models' parameters over different epochs, on the same underlying training set, is a key factor in boosting the MIA accuracy. However, for conducting their attacks, Nasr et al. make two rather strong assumptions. First, they assume white-box access to the target model; yet, most ML models, and especially the privacy-sensitive ones, are commonly accessible through APIs in a black-box manner. Second, they assume adversaries with access to a dataset, D', which partially overlaps with the target training set, D, without knowing, however, which data points are in $D' \cap D$. Finally, their MIA's performance drops as the number of participants in the collaborative learning system increases. For example, the accuracy of their passive attacker drops from 89.0% to 67.2% when going from 2 to 5 participants on CIFAR-100. As discussed in Sec. 2.4, tasks with many classes are more vulnerable to MIAs than tasks with fewer classes. Thus, this drop in performance might be even worse for datasets with fewer classes (e.g., CIFAR-10 & SVHN).

Leino et al. [36] propose white-box MIAs that operate without access to *any* of the target model's training data, thus relaxing Nasr et al.'s second assumption. Their work uncovers a more intimate understanding of how overfitting takes place in a model, which is not necessarily manifested in the model's output behaviour. In other words, the model is considered well-generalizable by the conventional overfitting detection methods (see Sec. A.1.1). In general, Leino et al. show that even if the target model is well-generalizable, a white-box adversary can infer the membership status of specific data records that the model has memorized, because this memorization is likely to show up in the way that the model uses features. Nonetheless, Leino et al.'s MIAs impose the following limitations. First, similar to Nasr et al. [50], they assume adversaries who can access the target model's internals, including its architecture and training algorithm. Second, their MIAs outperform previous white-box attacks (e.g., Nasr et al.) only by a very small margin.

Long et al.'s [44] black-box MIAs are inspired from Leino et al.'s statement that *if the adversary confidently identifies even one training point, then it is reasonable to say that a privacy violation occurred [36].* Following this direction, Long et al. show that even if a discriminative ML model is well-generalizable and achieves state-of-the-art performance, it *may* still contain *vulnerable data records (outliers)* that can be exploited with partial knowledge of

Table 1: A summary of the MIA papers found in the literature sorted in ascending chronological order. "-" means not specified or does not apply. "✔" (and "✗") symbol means that the respective adversary meets (or does not meet) each column's point. The highlighted rows represent MIAs that demonstrate *moderate* or *high* performance on well-generalizable targets. In addition, we use the following notation: white-box (WB), black-box (BB), fully black-box (FBB), sample-level (SL), user-level (UL). For more details on the inference levels as well as the attack settings see Sec. 2.1 and 2.3, respectively.

	Paper	Attack Setting	Target Environment		No. of shadow models	Agnostic regarding the target			Level of inference	Performance (low, moderate, high) on			Targeted instances	Utilized dataset(s)	
			Standalone	Collaborative		Model's internals	Data distribution	Sample's true class		Overfitted model	Well-generalizable model	Statistics or embeddings		Tabular & text dataset(s)	Image dataset(s)
2008	Homer et al. [22]	BB	Aggregate statistics (genomic data)	-	-	-	✗	-	UL	-	-	High	All	WTCCC, HapMap	-
2009	Wang et al. [76]	BB	Aggregate statistics (genomic data)	-	-	-	✗	-	UL	-	-	High	All	HapMap	-
2015	Dwork et al. [13]	BB	Aggregate statistics (distorted genomic data)	-	-	-	✗	-	UL	-	-	High	All	- (Theoretical analysis only)	
2016	Backes et al. [1]	BB	Aggregate statistics (microRNA expression data)	-	-	-	✗	-	UL	-	-	Moderate	All	GEO database: reference GSE61741	-
2017	Buscher et al. [4]	BB	Aggregate statistics (energy consumption data)	-	-	-	✗	✗	UL	-	-	2 users: High, >2 users: Low	All	Dataport, Redd, AMPds, ECO, UCI, GOVAU, UMASS	-
	Shokri et al. [64]	BB	Discriminative (MLP, CNN, MLaaS)	-	Multiple	✗	✗	✔	SL	High	Low	-	All	Purchases, Locations, Texas, Adult	MNIST, CIFAR-10/100
	Song et al. [67]	WB FBB	Discriminative (SVM, LR, ResNet, CNN)	-	0	✗ ✗	✔ ✔	✔ ✗	SL	- -	High High	-	Subsets of the training data	News, IMDB	CIFAR-10, LFW, FaceScrub
2018	Pyrgelis et al. [55]	BB	Aggregate statistics (location time-series)	-	-	-	✗	✗	UL	-	-	High	All	TFL, SFC	-
	Yeom et al. [79]	WB FBB	Discriminative (RR, DT, CNN)	-	1 0	✗ ✗	✗ ✗	✗ ✗	SL	High	Moderate	-	All	Eyedata, IWPC, Netflix	MNIST, CIFAR-10/100
	Hayes et al. [16]	WB BB BB	Generative (DCGAN, DCGAN+VAE, BEGAN)	-	1 1 0	✗ ✔ ✔	✗ ✔ ✗	✗ ✔ ✗	SL	High Low Moderate	-	-	All	-	LFW, CIFAR-10, Diabetic Retinopathy
2019	Melis et al. [47]	WB	-	Discriminative (VGG, RNN, CNN)	0	✗	✔	✔	SL	-	Low	-	All	Yelp-health, Yelp-author, CSI, FourSquare	-
	Sablayrolles et al. [59]	BB	Discriminative (ResNet, VGG)	-	0	✗	✗	✗	SL	High	Moderate	-	All	-	CIFAR-10, Imagenet
	Nasr et al. [50]	WB WB	Discriminative (LR, CNN)	Discriminative (LR, CNN)	Multiple 0	✗ ✔	✗ ✗	✗ ✗	SL	High High	Moderate Moderate	-	All	Purchase-100, Texas-100	CIFAR-100
	Salem et al. [60]	BB BB BB	Discriminative (LR, RF, MLP, CNN, MLaaS)	-	1 1 0	✔ ✔ ✔	✗ ✔ ✔	✔ ✔ ✔	SL	High Moderate Moderate	Low Low Low	-	All	Locations, Purchases, Adult, News	MNIST, CIFAR-10/100, Face
	Hilprecht et al. [20]	BB WB	Generative (Any) Generative (VAE)	-	1 0	✔ ✗	✗ ✗	✗ ✗	SL	Moderate High	Low Low	-	All	-	MNIST, Fashion-MNIST, CIFAR-10
	Song et al. [68]	FBB	Generative (LSTM, Seq2Seq)	-	Multiple	✗	✔	✗	UL	-	High	-	Text with rare words	Reddit, SATED, Dialogs, Wikitext-103	-
2020	Jayaraman et al. [27]	WB	Discriminative (MLP)	-	Multiple	✗	✗	✗	SL	High	-	-	All	Purchase-100, Texas-100, RCV1X	CIFAR-100
	He et al. [18]	WB BB	Discriminative (PSPNet, UperNet, Deeplab-v3+, DPC)	-	1	✗ ✔	✗ ✔	✗ ✔	SL	High	-	-	All	-	Cityscapes, BDD100K, Mapillary Vistas
	Song et al. [66]	WB	Discriminative (LSTM, Transformer, BERT, ALBERT)	-	-	✗	✗	✗	UL	-	-	-	All	Wikipedia articles, BookCorpus	-
	Chen et al. [7]	WB BB FBB	Generative (PGGAN, WGANGP, DCGAN, MEDGAN, VAEGAN)	-	0	✗ ✔ ✔	✗ ✗ ✗	✔ ✔ ✔	SL	High	Low	-	All	MIMIC-III, Instagram New-York,	CelebA
	Leino et al. [36]	WB	Discriminative (MLP, CNN)	-	1	✗	✔	✗	SL	High	Moderate	-	All	Adult, Diabetes, Cancer, Hepatitis, German Credit	MNIST, CIFAR-10/100, LFW
	Long et al. [44]	BB	Discriminative (MLP, CNN)	-	Multiple	✗	✗	✗	SL	-	High	-	Outliers	Adult, Cancer	MNIST
	Shadi et al. [56]	FBB	Discriminative (MLP, CNN)	-	1	✗	✗	✗	SL	High	Low	-	All	Purchase-100, Texas-100, Locations	MNIST, Fashion-MNIST, CH-MNIST, CIFAR-10/100
2021	Liu et al. [39]	BB BB BB BB	Image encoder (ResNet, VGG, CLIP)	-	1	✗ ✗ ✔ ✔	✗ ✔ ✗ ✔	✗ ✗ ✗ ✗	SL	Moderate High High High	- - - -	-	All	-	CIFAR-10, STL-10, Tiny-ImageNet
	Hui et al. [25]	BB BB WB WB	Discriminative (MLP, CNN)	-	0	✔ ✔ ✗ ✗	✔ ✗ ✔ ✗	✔ ✗ ✔ ✗	SL	High	-	-	All	Adult, Locations, Purchase-50, Texas	EyePACS, CH-MNIST, CIFAR-100, Birds-200
	Song et al. [69]	BB	Discriminative (MLP, CNN)	-	Multiple	✔	✗	✗	SL	Low	-	-	All	Purchase-100, Texas-100, Location-30	CIFAR-100, CH-MNIST, Car196
	Shafran et al. [62]	BB	Image translation/semantic segmentation (NVAE, Pix2PixHD, UperNet, HRNetV2)	-	0	✔	✗	✗	SL	High	Low	-	All	-	CelebA, Maps2sat, Cityscapes, CMP Facades, ADE20K
	Li et al. [38]	FBB FBB	Discriminative (CNN)	-	1 0	✗ ✔	✗ ✔	✔ ✔	SL	Moderate	-	-	All	-	CIFAR-10/100, Face, GTSRB
	Choquette-Choo et al. [8]	FBB	Discriminative (MLP, CNN, ResNet)	-	1	✗	✗	✗	SL	High	-	-	All	Texas-100, Purchase-100, Locations, Adult	MNIST, CIFAR-10/100
	Zhang et al. [83]	FBB	Discriminative (Item, LFM, NCF)	-	1	✗	✗	✗	UL	-	-	-	All	ADM, Lf-2k, Ml-1m	-
2022	Carlini et al. [5]	BB BB	Discriminative (WRN, VGG, ResNet, DenseNet, Inception-v3, MobileNet-v2), Generative (GPT-2)	- -	Multiple 0	✗ ✗	✗ ✗	✗ ✗	SL	High Moderate	Moderate Low	-	All	WikiText-103	CIFAR-10/100, ImageNet
	Ye et al. [78]	BB	Discriminative (AlexNet, VGG, WRN, CNN, MLP)	-	Multiple	✗	✗	✗	SL	High	-	-	All	Purchase-100	MNIST, CIFAR-10/100
	Liu et al. [42]	BB BB BB FBB	Discriminative (VGG, ResNet, WRN, MobileNet-v2)	-	Multiple	✗ ✗ ✔ ✗	✗ ✔ ✗ ✗	✗ ✗ ✗ ✗	SL	High Moderate Moderate Moderate	-	-	All	News, Purchase-100, Location-30	CIFAR-10/100, GTSRB, CINIC-10

specific instances from the target model's training set. In that sense, the authors argue that instead of proposing adversaries who indiscriminately attack all the records without regard to the cost of false positives or negatives, a more pragmatic approach would be to conduct MIAs on carefully selected vulnerable records. As a consequence of this vulnerable to MIAs record selection, the adversary minimizes false positives and boosts its precision. Note, however, that several techniques for reliably detecting and excluding vulnerable instances (i.e., outliers) have been proposed [2, 77, 80]. Thus, in case the target model's dataset is preprocessed, Long et al.'s MIAs *fall short*. Even if the target dataset is not preprocessed, the adversary can only identify a very small set of vulnerable instances. For example, Long et al. detect only 1 and 7 vulnerable records from the entire MNIST and Adult datasets, respectively.

For conducting their black-box MIAs, Carlini et al. [5] first train N shadow models on samples from the data distribution D, so that half of these models are trained on the target point (x, y), and half are not (they call these respectively IN and OUT models). Then, they fit two Gaussians to the confidences of IN and OUT models on (x, y), query the confidence of the target model on (x, y), and output a parametric likelihood-ratio test. However, when evaluating their attacks, Carlini et al. assume that the training sets of individual shadow models and the target model *partially overlap* –same as Nasr et al. Moreover, although black-box in nature, Carlini et al.'s MIAs require the knowledge of the target model's internals for constructing/training similar shadow models and achieving moderate success rates on well-generalizable targets. Finally, Carlini et al.'s MIAs are computationally expensive since, for each target record, they must train N different IN and OUT shadow models.

Song et al. [67] examine the case where a malicious ML provider supplies model-training code to the data holder (victim), without observing the training phase, and then obtains white- or fully black-box access to the trained model. The malicious code, when executed on the victim's sensitive data, produces models that are well-generalizable, yet, leak information about their training instances. In particular, it augments the training set with synthetic inputs whose labels encode information about the original training set. When the model is trained on the augmented dataset it becomes overfitted to the synthetic inputs. As a result, when the adversary submits one of these synthetic inputs, the model outputs the label that was associated with this input during training, thus leaking membership information. Song et al. show that using third-party code to train ML models on sensitive data is risky even if the code provider does not monitor the actual training phase. However, it is realistic to assume that highly sensitive datasets will be treated by experts in ML who will carefully inspect the model-training code before using it. Moreover, Song et al. can only infer the membership status of a small subset of the training set that the model memorized during training; if an adversary tries to memorize the whole training set, then this might impact the model's performance on the main task, and thus reveal that the model is overfitted.

Song et al. [68] propose black-box MIAs on ML models that generate natural-language text, such as word prediction and dialogue generation. Such models are at the core of popular online services and are often trained on personal data, such as users' messages, searches, chats, and comments. Inspired from their previous work [67], Song et al. analyse how text-generation models memorize word sequences since this can make them susceptible to MIAs. They find that such models overfit (memorize) text sequences containing *rare* words, thus, making those sequences vulnerable. Similar to their previous work, they show that the overfitting on these sequences does not appear in the model's testing accuracy, meaning that their target models are well-generalizable, rather than on the ranking of the candidate words that it generates. Nonetheless, Song et al.'s MIAs suffer from the following limitations: (a) they assume adversaries that have access to a subset of the victim's actual texts, (b) their MIAs can only succeed on text sequences containing rare words, and thus cannot target any user, and (c) their MIAs can be *eliminated* using Differential Privacy (DP) techniques for training the target model; DP guarantees that the influence of rare words is the same as all the other words in the training set.

Finally, the following general observations stem from our systematic analysis of the related literature:

- **Observation 1:** MIAs are easier on overfitted models rather than well-generalizable ones.
- **Observation 2:** All black-box MIAs achieving moderate or high success rates on well-generalizable targets require the knowledge of the target samples' true class.
- **Observation 3:** MIAs on well-generalizable models, although not effective on all target instances, perform well-above baseline on records which are highly influential to the model's parameters, such as outliers or out-of-distribution points.

3.2 Questions Remaining Unanswered

In this section, we identify a series of aspects that are yet to be explored. Below, we elaborate on these aspects and raise specific Research Questions (RQs) that we answer in the following sections.

First, the generalization of an ML model is directly related to its vulnerability against MIAs (the higher its generalization, the higher its robustness). In addition, the generalization of an ML model heavily depends on the utilized optimizer and regularization method(s). Nonetheless, *it is currently unclear whether specific generalization techniques result in models which are more vulnerable to MIAs than other models trained with different generalization techniques (RQ-1).*

Second, as shown in Table 1, most prior works are *not* agnostic regarding the target model's data distribution and the samples' true class. For example, some of them generate instances coming from the same distribution as the target model's dataset for which they know their membership status, for training shadow models [60, 64], and others assume access to a small subset of the actual target model's dataset, for demonstrating that even well-generalizable models may contain vulnerable records that can be exploited [44, 68]. Nonetheless, the increasing number of data breach incidents begs the question of *what is the actual benefit provided to potential adversaries in case they have access to a large portion of the target model's instances for which they know, not their true class, but their ground-truth membership status (RQ-2).* Besides, as Salem et al. [60] suggest, obtaining the instances' true class could be hard in certain cases (e.g., in biomedical settings).

Third, the structure and type of an ML model also contribute to its vulnerability against MIAs [64]. For example, Srivastava et al. [71] explain that deep ANNs' complexity (depth) increases the chances of overfitting, and thus potentially increasing their vulnerability to MIAs, since their capacity for memorizing training records

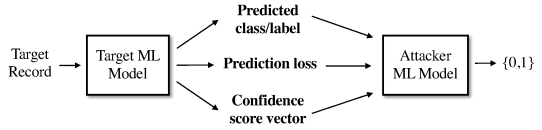


Figure 1: The black-box MIA cases as reported in [59]. Our paper focuses on the third case where the adversary feeds the target model with the target record and receives the confidence score vector as a response. Next, the adversary feeds the confidence score vector to the attacker model to infer whether the target record was included in the target model's training dataset (responds 0) or not (responds 1).

increases. However, *an experimental exploration of the aforementioned aspect (i.e., whether or not the number of layers facilitate MIAs and to what extent) is currently absent from the literature (RQ-3).*

In this paper, we answer the RQs raised above following an experimental approach. For the reasons mentioned in Sec. 2.3, we conduct our entire experimental analysis using black-box adversaries only.

4 EVALUATING COMMON ML GENERALIZATION TECHNIQUES

In this section, we investigate which learning settings may contribute to membership inference and to what extent. For exploring this problem we utilize ResNet-18 as the target model. Then, we vary each learning setting described below and observe its contribution to the MIA success rates. For each setting that we vary, we train a model for 100 epochs. Then, for ensuring that our models are well-generalizable, we measure the overfitting gap as the difference between the training and testing accuracies. As shown in Table 2, our target classifiers achieve high training/testing accuracies on CIFAR-10, while also being well-generalizable.

Similar to Long et al. [44], we take a numerical analysis approach and estimate the MIA success rate using the Monte Carlo method. In particular, for each experiment we randomly select the instances for composing the attacker model's training/testing subsets, and thus avoid any bias towards selecting a specific training/testing subset. We repeat this random (uniform) selection process 10 times and report the average attack success rate [4]. Finally, we utilize an MLP as the attacker model (see Fig. 8, Appx. A.1).

All the experiments reported in this section are conducted in the *optimal black-box attack setting*. That is, we utilize the black-box attack pipeline shown in Fig. 1 assuming adversaries that: (a) have access to *all* of the target model's train/test instances along with their ground-truth membership status, and (b) can query the target model *arbitrary* times for obtaining its confidence scores on each of those records. By doing so, we will expose any differences, in terms of privacy vulnerabilities, that the tested optimizers and regularization methods may have, against black-box adversaries with *strong background knowledge*. Note that there are many weaker black-box adversaries in the literature which are omitted since they do not change our conclusions.

Finally, in this section, we do *not* aim to demonstrate that increasing a model's generalization (e.g., by using regularization techniques) leads to lower MIA success rates than those achieved on an overfitted target –this is a well-known fact [60, 79]. Instead, we aim to *compare* the MIA success rates achieved on well-generalizable ML models trained with common optimizers and regularizers.

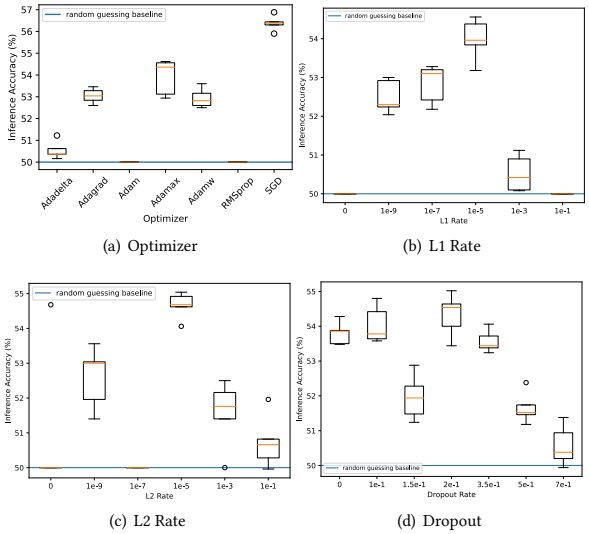
[4]For more details on the data splits see Sec. 5.

Table 2: ResNet-18 training/testing accuracies on CIFAR-10 for each optimization and regularization method used.

	Selection	Training Acc.	Testing Acc.
Optimizer	**SGD [32]**	94.23%	93.07%
	RMSprop [73]	89.72%	88.84%
	Adam [33]	90.78%	90.34%
	Adamax [33]	88.63%	88.49%
	Adadelta [81]	75.87%	74.78%
	Adagrad [12]	84.11%	83.03%
	AdamW [45]	90.23%	89.12%
*l*1 Rate	**0**	90.78%	90.34%
	1e-9	90.23%	88.96%
	1e-7	91.18%	88.07%
	1e-5	93.72%	90.74%
	1e-3	80.56%	77.47%
	1e-1	38.54%	35.23%
*l*2 Rate	**0**	90.78%	90.34%
	1e-9	90.67%	88.25%
	1e-7	92.43%	89.29%
	1e-5	93.76%	90.57%
	1e-3	88.23%	86.32%
	1e-1	53.89%	49.97%
Dropout Rate	**0**	90.78%	90.34%
	1e-1	94.25%	90.89%
	1.5e-1	90.23%	88.68%
	2e-1	93.42%	90.40%
	3.5e-1	90.21%	87.55%
	5e-1	85.39%	83.85%
	7e-1	75.94%	71.19%

Assessing Optimization Algorithms. A large number of ANNs' *optimizers* exist; each optimizer may converge to a different local optima, and thus a different set of weights' values. Optimization functions aim to minimize a specific notion of prediction error/loss (e.g., Mean Squared Error (MSE), $l1$ loss, or $l2$ loss –Eq. (3), Appx. A.1), for finding the global optima (i.e., the weights' values that result in the least possible prediction error). A robust optimizer should consider the facts that many local optima exist (i.e., higher prediction error points that our classifier may stuck in), and the prediction error surface has a complex morphology and thus the hyper-parameters of the optimizer (e.g., learning rate & decay factor), are of utmost importance. In this section, we conduct MIAs against ML models trained using the most popular optimizers, namely: (a) Stochastic Gradient Descent (SGD) [32], (b) RMSprop [73], (c) Adagrad [12], (d) Adadelta [81], (e) Adam [33], (f) Adamax [33], and (g) AdamW [45]. In particular, we deploy each of the aforementioned optimizers to train ResNet-18 on CIFAR-10. Next, we perform MIAs against each (trained) target ML model for concluding about the *information leakage level* of each optimization algorithm.

Figure 2(a) shows the MIA success rates achieved on ResNet-18 when trained with each particular optimizer. As shown, no significant difference, in terms of inference accuracy achieved for each optimizer, exists. In other words, all tested optimizers perform comparably well in terms of resisting against MIAs in the optimal black-box attack setting, while also delivering high-performing models (see Table 2). Thus, based on our experimental evidence, we conclude that the choice of the optimizer has *little impact* on the privacy robustness of well-generalizable models against black-box adversaries with strong background information.



(a) Optimizer (b) L1 Rate

(c) L2 Rate (d) Dropout

Figure 2: MIA success rates on ResNet-18 trained on CIFAR-10 using various optimization and regularization methods.

Assessing Regularization Methods. *$l1$ and $l2$ regularization methods* add a penalty parameter to the cost function that we essentially need to reduce. The key difference between $l1$ and $l2$ is the calculation of the penalty term. $l1$ adds the *absolute value of magnitude* of coefficient as the penalty term with a parameter λ to the cost function, whereas $l2$ adds the *squared magnitude* of coefficient as the penalty term with a parameter λ to the cost function. In both cases, λ is a hyper-parameter whose value is decided by the user. If λ is high it might prevent the model from learning (underfitting); if λ is close to 0 it may minimize the effect on the penalty term causing no regularization. Having the MSE as the cost function, $l1$ and $l2$ regularization terms are highlighted in equations (4) and (5) (Appx. A.1), respectively [5]. The target model's training/testing accuracies achieved for different $l1$ and $l2$ rates are shown in Table 2. Note that we utilize SGD for training the target models shown in Table 2 with different $l1$, $l2$ and dropout rates.

Figure 2(b) shows the average MIA success rates achieved for different $l1$ rates. As shown, no significant difference, in terms of MIA success rates, exists for each $l1$ rate; all average MIA success rates remain close to the random guessing baseline. In particular, even if $l1$ rate is 0, the average MIA success rate is 50%. In other words, a well-generalizable model causes our black-box MIAs to fail even if $l1$ is disabled. However, as seen in Table 2, selecting $l1$ rate $\geq$ 1e−3 largely affects the target model's performance. Thus, selecting $l1$ rate $\leq$ 1e−5, is considered enough for increasing both the target model's generalization and its robustness against MIAs.

The inference accuracies achieved for different $l2$ rates are shown in Figure 2(c). Again, no significant difference, in terms of MIA success rates, exists for each $l2$ rate; all average MIA success rates remain close to the random guessing baseline. Similar to $l1$ rate, even if $l2$ is 0, the average MIA success rate is 50%. In other words, a well-generalizable model causes our black-box MIAs to fail even if $l2$ is disabled. In addition, as shown in Table 2, selecting $l2$ rate $\geq$ 1e−3 largely affects the target model's performance. Thus, selecting $l2$ rate $\leq$ 1e−5, is adequate for increasing the target model's generalization as well as its robustness against MIAs.

Dropout causes the target ANN to randomly *drop out* units (neurons) by zeroing its weights' values [71]. This technique prevents the ANN from converging into a suboptimal solution, which is specialized only on a subset of the training set, due to the affection of the connections between the neighbouring neurons. Randomly dropping neurons will minimize each neuron's effect on its neighbours. Thus, dropout prevents overfitting by causing random neurons to detect the complex patterns hidden in the training data without being highly affected by their neighbours.

Figure 2(d) shows the average MIA success rates for different dropout rates. As shown, the MIA success rates remain close to the random guessing baseline, the highest being $\approx$ 54.5%, for each dropout rate. In fact, even if dropout is 0 (disabled), the average MIA success rate is $\approx$ 54%. Nonetheless, as shown in Table 2, selecting dropout $\geq$ 5e−1 significantly affects the performance of the target model. As a result, selecting a lower dropout rate (i.e., $\leq$ 3.5e−1) is considered enough for increasing both the target model's generalization and its robustness against MIAs.

ML Generalization Mechanisms vs. MIAs. Our findings suggest that the choice of the optimizer to be used for approximating the optimal solution as well as the regularization technique to be used for penalizing the model's weights has *no* discernible influence on the target model's robustness against MIAs. One can simply deploy an optimizer of their choice and set a relatively common value for the selected regularization technique. In particular, even if the selected regularization method is disabled, a well-generalizable (due to the optimizer) ML model causes black-box adversaries with strong background knowledge to fail on distinguishing the member from non-member instances. This demonstrates that not using any regularization technique does not necessarily lead to high MIA success rates. Furthermore, since we can train (well-generalizable) ML models which are robust against black-box MIAs with or without using a regularizer, we conclude that *MIA success rate is more closely related to the target model's generalization rather than to a particular regularization technique used.*

Another interesting observation stemming from our experimental evidence is the following. Even in the case where large regularization values lead to low performance ML models (that do not achieve state-of-the-art testing accuracies), which, however, have small generalization gap (<15%), the MIA success rates achieved from adversaries in the optimal black-box attack setting still remain close to the random guessing baseline. This fact demonstrates that *MIA success rates are not directly related to the performance of the target model rather than its generalization gap.*

We validate our conclusions by performing the same experiments using different image (CIFAR-100, SVHN), tabular (Adult, Surgical), and text datasets (IMDB), and observing similar results. However, due to space constraints, we placed those results in the appendix (see Figs. 9-13, Appx. A.2). Finally, note that exploring additional generalization techniques (e.g., batch/layer normalization, data augmentation), is an interesting direction for future research.

[5] We just use MSE for showing $l1$ and $l2$ regularization terms since it is one of the most popular loss functions. However, in our experiments, we utilize Cross-Entropy (CE) loss since it is a more natural choice when dealing with classification problems.

5 DATA LEAKAGE IMPLICATIONS

As Hui et al. [25] suggest, a powerful black-box MIA requires enough *labeled* output probability distributions of instances, included or not in the target model's training set, for learning the decision boundary of complex hyper-dimensional spaces. While this has been shown for overfitted models, it is currently unknown whether the same fact holds for well-generalizable ones. The ground-truth membership status of data records is commonly *unavailable* given only black-box access to the target model. Nonetheless, such information may be revealed in case of potential *data breaches*. In case of such incidents, one could explore if Hui et al.'s observation holds for well-generalizable target models as well.

Data breaches have been largely experienced even from high-profile web services, undermining their users' security and privacy [75]. Thus, it is realistic to assume that in case of such incidents a portion of MLaaS platforms' datasets may fall in attackers' hands. For each leaked data record its membership status may be included or not. For the time being, despite the several cases of data breach incidents, their contribution to MIAs *remains unexplored*. In this section, we shed light on this question by simulating the case where a portion of a sensitive dataset (in our case CIFAR-10) has been leaked by hackers or insiders. In such case, an attacker may deploy black-box MIAs and explore whether partial knowledge of the ground-truth membership information can facilitate MIAs, and likely reduce the resilience of well-generalizable models.

Threat Model. We consider both black-box attack scenarios defined by Hui et al., one which is strict and practical (black-box blind) and one which makes a strong assumption regarding the adversary's knowledge about the target model's dataset (black-box).

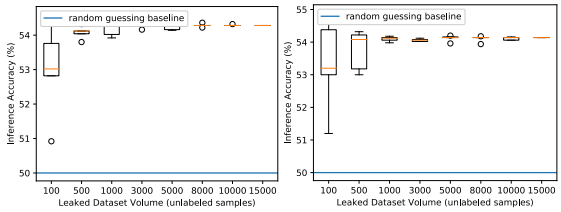
- *Black-box blind (most practical scenario):* The adversary can only query the target model and receive its confidence score vector as a response. However, the adversary does not have access to the target model's architecture, weights, or hyper-parameters, neither on the true class or the ground-truth membership status, of the queried instances.
- *Black-box (optimal black-box scenario):* This scenario is similar to the previous one, but assumes adversaries that have access to a portion of instances coming from the target model's dataset for which they know the members and non-members.

We define two adversaries (one for each scenario) that utilize off-the-shelf, and thus popular choices for potential attackers, ML models, namely MLP and k-means, for concluding about whether or not *unlabeled* (black-box blind) or *labeled* (black-box) data instances facilitate the successful deployment of MIAs. For both adversaries we train the attacker model having only *black-box* access to the target model (see Fig. 1). However, for the first adversary we assume that the input samples are unlabeled (i.e., they are not accompanied by the ground-truth membership status), whereas for the second adversary we assume that they are labeled. In any case, we avoid the need for shadow modelling since we assume adversaries that can query the actual target model *unlimited times* and receive, as a response, the confidence score vector for each leaked instance.

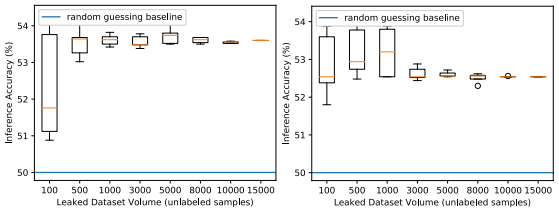
Experimental Setup. We split CIFAR-10 into a *train* and a *validation* subset containing 15,000 and 5,000 records, respectively. Both subsets are a split of 50%-50% member and non-member instances. Then, for examining the MIA success rates as the volume of leaked instances, l, increases, we randomly select l instances from the train subset. We repeat the random (uniform) selection process of the train subset's instances 10 times and report the average attack success rate. Finally, the inference accuracy results are reported on the *validation* subset. This evaluation methodology gives us a relatively accurate estimation of the success probability of our MIAs and complies with the state-of-the-art approaches found in the related literature [44, 60, 64].

Table 3: The target models' training/testing accuracies on CIFAR-10.

Architecture	Layers No.	Training Acc.	Testing Acc.
VGG [65]	11	100%	92.39%
	13	100%	94.14%
	16	99.61%	93.91%
	19	99.61%	93.79%
ResNet [17]	18	99.61%	93.03%
	34	99.61%	93.21%
	50	99.62%	93.64%
DenseNet [24]	121	99.22%	93.97%
	161	99.61%	94.07%
	169	99.62%	94.05%



(a) top-3 (b) top-5



(c) top-7 (d) top-10

Figure 3: Adversary 1's MIA success rates achieved on VGG-11 trained on CIFAR-10 using top-k confidence scores projection function ($k = 3, 5, 7, 10$). The inference accuracy is reported on the validation subset.

We target some of the most popular ANN architectures, namely VGG, ResNet, DenseNet and MLP, with various depths. These models have been trained on image, tabular and text datasets achieving state-of-the-art performance. In addition, we utilize early-stopping mechanism for avoiding overfitting. For brevity, in this section, we focus on VGG-11 trained on CIFAR-10 (see Table 3 for the training/testing accuracies) since the collected results are similar when using all other architecture-dataset combinations and due to space constraints. However, for the interested reader, we still provide those results in the appendix (see Appx. A.3).

Hui et al.'s black-box adversaries relax the assumptions that: (a) the target and attacker models have to share the same structure, and (b) the adversary knows about the target model's internals

(e.g., its architecture, gradients, loss function and hidden layers' activations), which are rather strong (especially when targeting privacy-sensitive models). In particular, we attack a CNN-based image recognition model using a statistical-based model, i.e., k-means (adversary 1), and a fully-connected MLP (adversary 2). Both MLP and k-means have been utilized in similar research achieving well-above baseline MIA success rates on overfitted targets [25, 60, 64]. Finally, for conducting our attacks, we utilize the probability score projection functions defined by Hui et al. The idea behind projection functions is that the ranking of values in different classes, rather than the actual label of the class, determines the membership. The projection functions used are: (a) top-k confidence scores for $k = 3, 5, 7, 10$ (adversary 1) and (b) top-k confidence scores + ground-truth membership status for $k = 3, 5, 7, 10$ (adversary 2).

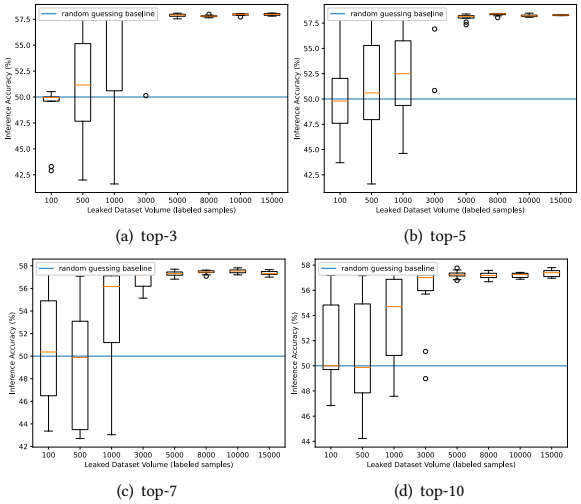
Black-box Blind Scenario – Adversary 1. *Operation.* For the first adversary, we assume *no* leakage of the ground-truth membership status of each instance. Instead, the attacker tries to distinguish between member and non-member records by just considering their respective confidence score vector derived from the target model. As a result, this adversary maps directly to the black-box blind setting which, according to [25], is the most strict and practical as it does not rely on any private information which might not be available to a potential attacker. Conducting MIAs in such an *agnostic* setting allows us to explore the *security* and *privacy vulnerability boundaries* of *well-generalizable* models.

The attacker model, namely k-means [6], clusters the confidence score vectors, derived from the last (output) layer of the target model, into two different groups, namely member and non-member records. In other words, this adversary receives as input the target model's confidence scores and decides about whether the input sample that yielded those confidence scores was included in the target model's training set or not. k-means exploits the knowledge hidden in the score vectors' feature-space by calculating a notion of similarity (i.e., Euclidean distance) and classifies each given input sample to one of the two predefined clusters with a certain level of confidence. Initially, we deploy k-means for use *without* training it at all; the training happens *on the fly* as the adversary feeds it with confidence score vectors derived from the target model.

Results. We examine this adversary's performance with respect to the number of unlabeled instances for estimating the minimum amount of input samples required for achieving high-enough MIA success rates. The attack success rates for different volumes of input samples are depicted in Figure 3. As shown, the achieved MIA success rates remain close to the random guessing baseline (50%) in all cases. In addition, the average MIA success rates remain approximately the same as we keep increasing the volume of input samples. Overall, adversary 1 achieves 53.6% MIA success rate on average, the highest being 54.3%, when targeting VGG-11. We validate our conclusions by conducting the same experiments on different (and deeper) architectures (ResNet, DenseNet, MLP) trained on datasets from various domains (CIFAR-100, SVHN, Adult, Surgical, IMDB), and observing similar results. However, due to space constraints, we placed those results in the appendix (see Appx. A.3).

In contrast, Hui et al. achieve well-above baseline MIA success rates using the same clustering algorithm (k-means), but attacking,

[6]For the implementation of k-means we utilize Scikit-learn (version 0.21.2) [53].



(a) top-3 (b) top-5

(c) top-7 (d) top-10

Figure 4: Adversary 2's MIA success rates achieved on VGG-11 trained on CIFAR-10 using top-k confidence scores + ground-truth projection function ($k = 3, 5, 7, 10$). The inference accuracy is reported on the validation subset.

however, *overfitted* models. Thus, based on our experimental evidence, we conclude that well-generalizable models are significantly more robust than overfitted models, against adversaries that only exploit the target instances' confidence score vectors.

Black-box Scenario – Adversary 2. *Operation.* For the second adversary, we assume that the ground-truth membership status for each instance included in the target dataset has been leaked. Thus, in this case, the adversary has an *extra* bit of information, which is whether or not the confidence score vector was yielded by a member or a non-member instance. Having such *strong auxiliary knowledge*, a potential adversary can now utilize supervised learning algorithms for constructing an approximation function $F : X \rightarrow Y$ that maps each given input $x \in X$ (i.e., the confidence score vector), to each target output class $y \in Y$ (i.e., member or non-member). This adversary maps directly to the black-box attack setting described by Hui et al.

The operation of this adversary is identical to the previous one, the only difference being that the confidence score vectors, derived from the last (output) layer of the target model, are now accompanied by their ground-truth membership status. Essentially, this adversary exploits a valuable piece of (sensitive) extra information which: (a) might be used to unveil any correlations between the confidence score vector and the target record's membership status that may exist in the latent space, and (b) is *directly* related to the adversary's goal, that is, to infer the membership status of each target record. For doing so, the adversary has to deploy a more sensible for the problem ML model, which in our case is the MLP shown in Fig. 8 (Appx. A.1).

Results. We examine this adversary's performance with respect to the number of *labeled* instances. The attack success rates for different volumes of labeled input samples are depicted in Fig. 4. Surprisingly, even in the case where adversaries have access to the ground-truth membership status of a large portion of the target

model's instances, they cannot achieve more than 55.43% MIA success rate on average, the highest MIA success rate being 58%, when targeting VGG-11. Again, the average MIA success rates remain approximately the same as we keep increasing the volume of training instances-labels (ground-truth membership status) pairs. The collected results suggest that even if black-box adversaries possess such a valuable piece of information, they still *cannot* achieve well-above baseline MIA success rates when targeting well-generalizable models. We validate our conclusions by conducting the same experiments on different (and deeper) architectures (ResNet, DenseNet, MLP) trained on datasets from various domains (CIFAR-100, SVHN, Adult, Surgical, IMDB), and observing similar results. However, due to space constraints, we placed those results in the appendix (see Figs. 14-19, Appx. A.3).

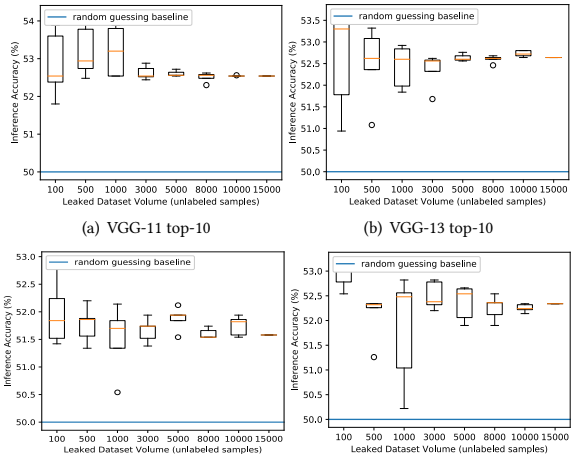
In contrast, some existing MIAs on *overfitted* targets achieve well-above baseline MIA success rates, using the same classification model (MLP), if such auxiliary knowledge is available to the attacker [60]. Thus, based on our experimental evidence, we conclude the following: (a) well-generalizable models are significantly more robust than overfitted models, even against adversaries that possess partial ground-truth membership information, (b) Hui et al.'s [25] observation (i.e., black-box MIA performance is analogous to the number of labeled instances available to the adversary) *does not* apply to well-generalizable targets.

Exploring ANNs' Depth Factor. An important factor concerning the performance of state-of-the-art ANNs is the number of layers. In general, traversing to deep ANNs, with a large number of layers, increases their prediction/classification accuracy compared to shallower ones [3]. However, increasing an ANN's complexity increases its memorization capabilities as well (necessary evil) [71]. As shown in Sec. 3.1.2, the memorization capabilities of (even well-generalizable) models may increase the vulnerability of certain instances, commonly the highly influential points, to MIAs. Thus, since increasing an ANN's layers increases, at the same time, its memorization capabilities, it might be the case that deep ANNs are more vulnerable to MIAs than shallower ones. In this section, we explore whether deep ANNs' memorization capabilities make them more vulnerable to black-box MIAs compared to shallower ones (or the opposite). For doing so, we deploy MIAs, in both black-box and black-box blind scenarios, on VGG models with various depths.

Figures 5 & 6 show the attack success rates achieved from adversary 1 (black-box blind) and 2 (black-box), respectively, using top-k confidence scores (adversary 1) + ground-truth membership status (adversary 2) projection functions with $k = 10$. As shown, the MIA success rates remain close to the random guessing baseline for all tested depths and for both attack scenarios. Thus, based on our experimental evidence, we conclude that deep ANN architectures are *not* more vulnerable to our black-box MIAs compared to shallower ones or the opposite. We validate our conclusions by also targeting DenseNet, ResNet and MLP architectures (with various depths) trained on datasets from different domains, and observing similar results. However, due to space constraints, we placed those results in the appendix (see Figs. 14-19, Appx. A.3).

6 DISCUSSION

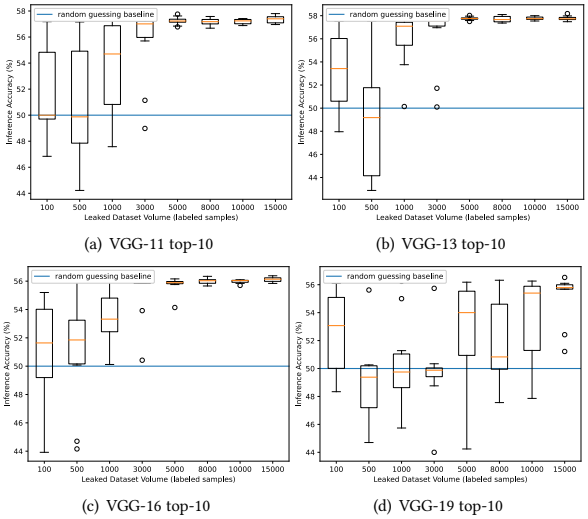
Why MIAs on Overfitted Targets Succeed? As shown in Table 1, the majority of prior research succeeds only on overfitted targets.



(a) VGG-11 top-10 (b) VGG-13 top-10

(c) VGG-16 top-10 (d) VGG-19 top-10

Figure 5: Adversary 1's MIA success rates achieved on VGG (with different number of layers) trained on CIFAR-10.



(a) VGG-11 top-10 (b) VGG-13 top-10

(c) VGG-16 top-10 (d) VGG-19 top-10

Figure 6: Adversary 2's MIA success rates achieved on VGG (with different number of layers) trained on CIFAR-10.

For acquiring such models the authors *intentionally* train their models for a large fixed number of epochs until a significant gap between the training and testing accuracies is presented. As Shokri et al. [64] explain, MIAs exploit the fact that ML models often behave differently on the data that they were trained on versus the data that they "see" for the first time. Overfitting the target model contributes to the aforementioned aspect (i.e., differentiating the decision behaviour for member and non-member samples), thus enabling MIAs [39]. However, even a *naive* MIA, which predicts a point is a member if and only if it was classified correctly [79], is highly effective on models that overfit to a large degree [36].

In contrast, well-generalizable models behave similarly on member and non-member samples, thus causing the majority of previous

MIAs to fail. As Rezaei et al. [58] suggest, for well-generalizable models, the only considerable difference appears between correctly classified samples and misclassified samples, not between members and non-members. However, as shown in our experimental analysis in Sec. 5, if the adversary does not have access to the true class of each training instance, even low performing ML models, but with a small gap between training and testing accuracies (< 15%), are robust against black-box adversaries with access to the ground-truth membership status of a large volume of leaked instances.

Note that we do not suggest, in any case, that training well-generalizable models is enough to mitigate any black-box MIA. In fact, black-box (or even fully black-box) MIAs achieving well-above baseline success rates on well-generalizable targets do exist (Sec. 3). Contrary, as the title suggests, we aim to demonstrate that *MIAs are harder than we previously thought* under reasonable assumptions.

Limitations and Future Work. Relevant works showed that overfitting, although sufficient, is *not* a necessary condition for enabling MIAs [50, 60, 64]. Below, we discuss other factors affecting MIAs and provide directions for future research.

Outliers, Highly Influential Points. Feldman and Zhang [14] show that not all training samples are equal; some samples (outliers) have an outsized effect on an ML model's parameters when inserted into its training set, compared to other (inlier) samples. As shown in our systematic analysis in Sec. 3.1, well-generalizable models can memorize (overfit) such highly influential points, thus making them vulnerable to MIAs (low-hanging fruits) [7, 8, 20, 44, 67, 68]. However, as Carlini et al. [6] explain, the unintended memorization of such instances is a persistent, hard-to-avoid issue that can have serious consequences. This aspect is further demonstrated in our experimental analysis in Sec. 5, where small deviations from the baseline exist, designating that specific instances are more vulnerable to MIAs than others. We attribute these deviations to specific high risk/vulnerable records as suggested by Long et al. [44] & Carlini et al. [5]. Note that one can utilize approaches to measure the risk of individual samples to verify this hypothesis [40]. However, since in this paper our main focus is assessing the MIA risk with respect to an adversary who *indiscriminately* attacks all the records in the target dataset, presenting such a detailed analysis in regards to specific vulnerable to MIA records, although indeed interesting, is considered as future work.

Kulynych et al. [35] demonstrate the phenomenon of disparate vulnerability against MIAs, that is, the unequal success rate of MIAs against different population subgroups. The authors explain that the vulnerability to MIAs arises when the distribution of a model's property (e.g., its loss, or outputs/logits) is different for member and non-member samples. In addition, they argue that preventing disparate vulnerability requires either: (a) increasing the complexity of the learning problem to ensure distributional generalization, or (b) using a differentially private training algorithm with the associated drop in performance. Much like the work of Kulynych et al., additional effort is required towards detecting (and defending) subgroups of the population which are vulnerable to MIAs.

Target Model's Type. The structure and type of the target ML model also contribute to its vulnerability against MIAs [64]. For example, as Truex et al. [74] show, a Naive Bayes model is much more resilient to MIAs than a decision tree and therefore may be the preferred model type for a particular ML service. Generally speaking, a target model whose decision boundary is unlikely to be drastically impacted by a particular instance will be more resilient to MIAs [74]. Nonetheless, a systematic analysis in regards to the resistance of different types of ML models and algorithms against state-of-the-art MIAs is currently absent from the literature.

In this paper, we focus our experimental analysis on ML models used for classification. Although demonstrating that the same observations apply to ML models trained on different tasks other than classification (e.g., generative models for text generation/image synthesis or embedding models) is beyond the scope of this paper, we consider this angle as an interesting direction for future research.

Attributing the Robustness of Well-generalizable Models to Specific Parameters/Settings. Despite that significant effort has been given on analysing the connections between overfitting and MIA success rate, there is a lack of deeper understanding of *why* well-generalizable models are much more resilient to black-box MIAs compared to overfitted ones, and how this knowledge can be potentially exploited to carry out successful MIAs against such models. In this paper, we follow an experimental approach, but further theoretical work is needed for having complete understanding of MIAs in the context of well-generalizable targets.

Providing theoretical evidence that well-generalizable models' robustness against MIAs is related to specific parameters/settings is one of the major directions for future work in this field. Having such knowledge, the scientific community could then introduce novel MIAs that achieve well-above baseline success rates on any record included in a well-generalizable model's dataset. In addition, the vulnerabilities of well-generalizable models (e.g., the case of outliers/highly influential points) could be further explored.

7 DEFENSES

Exploring a sizable portion of the attack space and defining the parameters/settings that impact MIAs' success rates is crucial for developing generally applicable defenses. In that sense, in this paper, we performed a systematic analysis of several parameters that may affect MIAs on well-generalizable targets. For the time being, none of the proposed defense strategies offers acceptable guarantees without sacrificing the target model's performance or the usefulness (utility) of the information provided.

Despite showing that well-generalizable models cause black-box MIAs with strong background information to fail (Sec. 5), below, we discuss defense strategies that will further enhance their resistance against superior adversaries that may follow a dramatically different attacking strategy. However, evaluating those defenses is out of scope, since our main goal in this paper is to systematically explore the different parameters that may make MIAs more effective without touching on the various choices a defender has access to per attack case. Thus, we provide this short review of possible defenses only for completeness and it is not by any means a thorough look at how these defenses work in each setting.

Differential Privacy. On the one hand, DP guarantees the low influence of each training instance, but on the other hand, it *significantly* reduces the performance of an ML model [29]. In fact, current DP mechanisms rarely offer acceptable utility-privacy trade-offs for complex learning tasks [27, 28]. Moreover, it was shown that even differentially private models are susceptible to sophisticated MIAs [55, 57]. In general, further research is necessary to understand

which DP mechanisms and under what settings are able to provide viable protection to MIAs without forfeiting the model's utility [74].

Adversarial Regularization. Nasr et al. [49] introduced a mechanism, namely adversarial regularization, to train models with membership privacy, which ensures indistinguishability between the predictions of a model on member and non-member samples. However, recent approaches managed to circumvent adversarial regularization and successfully deliver their attacks [8, 25, 69]. In addition, adversarial regularization is no better than early-stopping, which is a much simpler and computationally inexpensive strategy [69].

Li et al. [37] observe that a model's vulnerability against MIAs is tightly related to its generalization gap. Thus, similarly to Nasr et al. [49], they introduce a new regularizer to the training loss function in order to shrink this gap. In particular, the authors' regularizer penalizes the situation where member instances have significantly different output distributions from non-member ones.

Noise Adding. Adding random noise on the target model's confidence scores affects MIAs' performance [28]. Recently, Jia et al. [28] proposed a technique with formal utility-loss guarantees against black-box MIAs, namely MemGuard. Despite that MemGuard can be applied on any target model, the authors evaluate their defense's performance *only* on overfitted targets rather than well-generalizable ones. Moreover, MIAs that bypass the security provided by MemGuard were proposed [8, 25]. As Song et al. [69] suggest, MemGuard is *not* as effective as previously reported and lacks important evaluation against well-generalizable targets. Finally, the addition of random noise on the target model's confidence scores may undermine the utility of the results provided. More importantly, this alteration may have *serious consequences*, especially for models utilized in the healthcare field.

Restrict the Output Vector to Top-k Classes. The most straightforward defense strategy is to restrict the target model to only return the top-k confidence scores rather than the full vector with all classes [64]. Selecting only the top-k classes significantly limits the exploitable by adversaries information, and thus their overall performance. Generally speaking, the smaller the k is, the less information the model leaks [64]. Ideally, the model has to return only the index of the most likely class without even reporting its probability. Nonetheless, this defense strategy suffers from the following limitations: (a) similar to DP, it *degrades* the usefulness of the information provided by the model, and (b) it *cannot* protect against recent fully black-box MIAs that need access only to the target model's predicted (discrete) class [8, 38, 83].

8 RELATED WORK

Liu et al. [41] present an analysis of the risks caused by different inference attacks (e.g., various scenarios they can be applied to, common factors that influence their performance, or any relationship among them) and the effectiveness of defense techniques. They focus on four attacks, namely membership inference, model inversion [10], attribute inference, and model stealing, and establish a threat model taxonomy. Their experimental evaluation shows that: (a) the complexity of the training dataset largely affects the attacks' performance, and (b) the effectiveness of model stealing and membership inference are negatively correlated.

Papernot et al. [52] introduce a unifying threat model to allow structured reasoning about the security and privacy of systems that incorporate ML. In contrast to this paper, their threat model considers the entire data pipeline, of which ML is a component, instead of ML models in isolation. The authors provide an analysis regarding adversarial sample generation attacks [9] in various settings (e.g., white-box and black-box) and report the recent progress towards training robust, private, and accountable ML models.

Song et al. [69] perform an investigation in regards to why certain samples are more vulnerable to MIAs compared to other samples, including correlations with model properties, such as model sensitivity, generalization error, and feature embeddings. The authors stress the importance of a systematic evaluation of privacy risks of ML models, which is an angle that our work attempts to explore.

Song et al. [70] introduce MIAs that exploit structural properties of robust models on adversarially perturbed data. In particular, the authors show that compared to the natural training (undefended) approach, adversarial defense methods can *increase* the target model's risk against MIAs. Thus, in this paper, we focus our analysis solely on naturally trained ML models.

Hu et al. [23] present a comprehensive survey on MIAs and defenses, while also discussing their pros and cons. Based on the limitations identified, they provide promising future research directions. Contrary to Hu et al.'s work, our paper: (a) presents a systematization of knowledge that evaluates, systematizes, and contextualizes existing knowledge on MIAs, and thus provides useful insights that could not be obtained by simply reading each of the individual papers (or a survey paper that outlines such works), and (b) follows an experimental-based approach for examining specific parameters/settings that may affect black-box MIA success rates on well-generalizable targets.

Hyeong et al. [26] present an empirical analysis of MIAs in the context of tabular data synthesis models. According to the authors, the majority of prior research on MIAs that target generative models has concentrated in the image domain. Thus, the risk of tabular data synthesis models against MIAs is largely unexplored. Tabular data, however, are among the most widely used data types and frequently include sensitive or private information. The authors use Chen et al.'s [7] MIAs to show how tabular data synthesis models can be seriously jeopardized.

9 CONCLUSION

In this paper, we presented a systematization of knowledge for MIAs found in the literature. Based on our analysis, we raised specific questions that still remain unanswered. Next, we answered those questions in the following order. First, we compared the MIA success rates achieved on ML models trained with popular ML generalization techniques. Our results suggest that all tested generalization techniques demonstrate comparable robustness against adversaries in the optimal black-box attack setting. Second, we explored the contribution of potential data leaks to successful black-box MIAs. The collected results suggest that even if black-box adversaries possess partial ground-truth membership information, they still cannot achieve well-above baseline MIA success rates when targeting well-generalizable models. Third, we examined whether or not, and to what extent, the depth of ANNs facilitates black-box MIAs. In our experimental analysis we showed that this is not the case. Observing all the results reported in this paper, we conclude that MIAs are harder than we previously thought under reasonable assumptions.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback. We also thank the shepherd for helping us in improving the final version of this paper. This work was supported by the European Union's Horizon 2020 and Horizon Europe research and innovation programmes under grant agreements No. 101083594 (CyberSecPro), No. 101070599 (SecOPERA) and No. 101007673 (RESPECT).

REFERENCES

[1] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. 2016. Membership privacy in MicroRNA-based studies. In *CCS*. 319–330.
[2] Irad Ben-Gal. 2005. Outlier detection. In *Data Min. Knowl. Discov.* 131–146.
[3] Vitoantonio Bevilacqua, Antonio Brunetti, Andrea Guerriero, Gianpaolo Francesco Trotta, Michele Telegrafo, and Marco Moschetta. 2019. A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. In *Cogn. Syst. Res.*, Vol. 53. 3–19.
[4] Niklas Buescher, Spyros Boukoros, Stefan Bauregger, and Stefan Katzenbeisser. 2017. Two Is Not Enough: Privacy Assessment of Aggregation Schemes in Smart Metering. In *PETS*. 198–214.
[5] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022. Membership Inference Attacks From First Principles. In *S&P*. 1897–1914.
[6] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*. 267–284.
[7] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *CCS*. 343–362.
[8] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *ICML*. 1964–1974.
[9] Antreas Dionysiou and Elias Athanasopoulos. 2021. Unicode Evil: Evading NLP Systems Using Visual Similarities of Text Characters. In *AISEC*. 1–12.
[10] Antreas Dionysiou, Vassilis Vassiliades, and Elias Athanasopoulos. 2023. Exploring Model Inversion Attacks in the Black-box Setting. In *PETS*. 190–206.
[11] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[12] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. In *J. Mach. Learn. Res.*, Vol. 12. 2121–2159.
[13] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. 2015. Robust Traceability from Trace Amounts. In *FOCS*. 650–669.
[14] Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. In *NeurIPS*, Vol. 33. 2881–2891.
[15] Robert Hackett. 2017. Yahoo raises breach estimate to full 3 billion accounts, by far biggest known. (2017). https://fortune.com/2017/10/03/yahoo-breach-mail/
[16] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership inference attacks against generative models. In *PETS*. 133–152.
[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
[18] Yang He, Shadi Rahimian, Bernt Schiele, and Mario Fritz. 2020. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation. In *ECCV*. 519–535.
[19] Patrick Heim. 2016. Resetting passwords to keep your files safe. (August 2016). https://blog.dropbox.com/topics/company/resetting-passwords-to-keep-your-files-safe
[20] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models.. In *PETS*. 232–249.
[21] Geoffrey E Hinton, Alexander Krizhevsky, Ilya Sutskever, and Nitish Srivastva. 2016. System and method for addressing overfitting in a neural network. US Patent 9,406,017.
[22] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. In *PLoS Genetics*, Vol. 4. 1–9.
[23] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2021. Membership inference attacks on machine learning: A survey. In *CSUR*.
[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*. 4700–4708.
[25] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. 2021. Practical Blind Membership Inference Attack via Differential Comparisons. In *NDSS*.
[26] Jihyeon Hyeong, Jayoung Kim, Noseong Park, and Sushil Jajodia. 2022. An Empirical Study on the Membership Inference Attack against Tabular Data Synthesis Models. In *CIKM*. 4064–4068.
[27] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. 2021. Revisiting Membership Inference Under Realistic Assumptions. In *PETS*. 348–368.
[28] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *CCS*. 259–274.
[29] Aaron Johnson and Vitaly Shmatikov. 2013. Privacy-preserving data exploration in genome-wide association studies. In *SIGKDD*. 1079–1087.
[30] Poul-Henning Kamp, P Godefroid, M Levin, D Molnar, P McKenzie, R Stapleton-Gray, B Woodcock, and G Neville-Neil. 2012. LinkedIn password leak: salt their hide.. In *ACM Queue*, Vol. 10. 20.
[31] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2020. On the effectiveness of regularization against membership inference attacks. *arXiv preprint arXiv:2006.05336* (2020).
[32] Jack Kiefer, Jacob Wolfowitz, et al. 1952. Stochastic estimation of the maximum of a regression function. In *Ann. Math. Stat.*, Vol. 23. 462–466.
[33] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[34] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. *Learning multiple layers of features from tiny images*. Technical Report. Citeseer.
[35] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. 2022. Disparate Vulnerability to Membership Inference Attacks. In *PETS*. 460–480.
[36] Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *USENIX Security*. 1605–1622.
[37] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2021. Membership inference attacks and defenses in classification models. In *CODASPY*. 5–16.
[38] Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *CCS*. 880–895.
[39] Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. 2021. EncoderMI: Membership inference against pre-trained encoders in contrastive learning. In *CCS*. 2081–2095.
[40] Xiyang Liu, Yixi Xu, Shruti Tople, Sumit Mukherjee, and Juan Lavista Ferres. 2020. Mace: A flexible framework for membership privacy estimation in generative models. *arXiv preprint arXiv:2009.05683* (2020).
[41] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. 2021. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. *arXiv preprint arXiv:2102.02551* (2021).
[42] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership Inference Attacks by Exploiting Loss Trajectory. In *CCS*. 2085–2098.
[43] Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. 2017. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136* (2017).
[44] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen. 2020. A Pragmatic Approach to Membership Inferences on Machine Learning Models. In *EuroS&P*. 521–534.
[45] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
[46] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL HLT*. 142–150.
[47] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting Unintended Feature Leakage in Collaborative Learning. In *S&P*. 691–706.
[48] Fan Mo, Ali Shahin Shamsabadi, Kleomenis Katevas, Andrea Cavallaro, and Hamed Haddadi. 2019. Towards characterizing and limiting information exposure in DNN layers. *arXiv preprint arXiv:1907.06034* (2019).
[49] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine Learning with Membership Privacy Using Adversarial Regularization. In *CCS* (Toronto, Canada). 634–646.
[50] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *S&P*. 739–753.
[51] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop*.
[52] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. 2018. Sok: Security and privacy in machine learning. In *EuroS&P*. 399–414.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine

Learning in Python. In *J. Mach. Learn. Res.*, Vol. 12. 2825–2830.
[54] Lutz Prechelt. 1998. Early stopping-but when?. In *Neural Networks: Tricks of the trade*. 55–69.
[55] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2018. Knock Knock, Who's There? Membership Inference on Aggregate Location Data. In *NDSS*.
[56] Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. 2020. Sampling attacks: Amplification of membership inference attacks by repeated queries. *arXiv preprint arXiv:2009.00395* (2020).
[57] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. 2018. Membership Inference Attack against Differentially Private Deep Learning Model. In *Trans. Data Priv.*, Vol. 11. 61–79.
[58] Shahbaz Rezaei and Xin Liu. 2021. On the difficulty of membership inference attacks. In *CVPR*. 7892–7900.
[59] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*. 5558–5567.
[60] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *NDSS*.
[61] Daniel I Sessler, Andrea Kurz, Leif Saager, and Jarrod E Dalton. 2011. Operation timing and 30-day mortality after elective general surgery. *AACRAT* 113, 6 (2011), 1423–1428.
[62] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. 2021. Membership Inference Attacks are Easier on Difficult Problems. In *ICCV*. 14820–14829.
[63] Virat Shejwalkar and Amir Houmansadr. 2021. Membership privacy for machine learning models through knowledge transfer. In *AAAI*. 9549–9557.
[64] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *S&P*. 3–18.
[65] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[66] Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *CCS*. 377–390.
[67] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine learning models that remember too much. In *CCS*. 587–601.
[68] Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *SIGKDD*. 196–206.
[69] Liwei Song and Prateek Mittal. 2021. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *USENIX Security*. 2615–2632.
[70] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy risks of securing machine learning models against adversarial examples. In *CCS*. 241–257.
[71] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. In *J. Mach. Learn. Res.*, Vol. 15. 1929–1958.
[72] Jasper Tan, Blake Mason, Hamid Javadi, and Richard Baraniuk. 2022. Parameters or Privacy: A Provable Tradeoff Between Overparameterization and Membership Inference. In *NeurIPS*.
[73] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012), 26–31.
[74] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Demystifying membership inference attacks in machine learning as a service. *IEEE Trans. Serv. Comput.* (2019).
[75] Ding Wang, Haibo Cheng, Ping Wang, Jeff Yan, and Xinyi Huang. 2018. A Security Analysis of Honeywords. In *NDSS*. 1–16.
[76] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 2009. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study. In *CCS*. 534–544.
[77] Sylvain Weber. 2010. bacon: An effective way to detect outliers in multivariate data using Stata (and Mata). *The Stata Journal* 10, 3 (2010), 331–338.
[78] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. In *CCS*. 3093–3106.
[79] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*. 268–282.
[80] Dantong Yu, Gholamhosein Sheikholeslami, and Aidong Zhang. 2002. Findout: finding outliers in very large datasets. In *KAIS*. 387–412.
[81] Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).
[82] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. In *Commun. ACM*. 107–115.
[83] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership inference attacks against recommender systems. In *CCS*. 864–879.

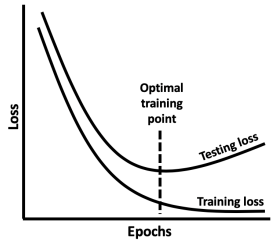


Figure 7: The training-testing loss vs. epoch graph for an *overfitted* model. For ensuring the maximum generalization of this model we had to stop the training process at the *optimal training point*. Beyond that point, the model becomes less generalizable and begins to memorize the labels of the training data at the expense of its ability to generalize.

A APPENDIX

The appendix is composed of three sections. First, in Sec. A.1, we provide supplementary material that is mentioned in this paper's main body. Second, in Sec. A.2, we provide additional results using different architecture-dataset combinations than those utilized in Sec. 4. Third, in Sec. A.3, we present the MIA success rates achieved from adversary 2 on different datasets and ANNs with various depths. Note that the information provided in each section is not meant to be read in a row.

A.1 Supplementary Material

A.1.1 Detecting and Tackling Overfitting. In supervised learning, there are two ways for detecting overfitting: (a) observing the difference between the training and testing accuracies, and (b) plotting the training-testing loss vs. epoch graph. Using the former, one can detect overfitting if the training and testing accuracies differ significantly, usually more than 10–15%, the testing accuracy being the lower one. Using the latter, one can detect overfitting by comparing the training-testing loss vs. epoch graph with the typical overfitted model's graph morphology shown in Fig. 7. As shown, if the testing loss starts to rise at a certain point while the training loss keeps decreasing then this is a clear evidence that the model started to overfit on the training data.

As a response, the scientific community came up with a number of techniques and mechanisms for preventing models from becoming overfitted to their training data, such as the early stopping [54] and regularization methods (e.g., $l1$, $l2$, and dropout) [21].

A.1.2 Attacker Model. Figure 8 shows adversary 2's MLP. The attacker model receives as input the confidence score vector derived from the target model when queried with a target record and responds 0 or 1 depending on whether or not the target record was included in the target model's training dataset. All layers utilize ReLU activation function (Eq. 1) except from the output layer that utilizes Softmax (Eq. 2).

A.1.3 Equations.

$$ReLU : f(x) = max(0, x) \tag{1}$$

$$softmax(z_i) = \frac{exp(z_i)}{\sum_{i=0}^{k-1}(exp(z_i))} \quad for\ i = 1, ..., K\ and\ z = (z_1, ..., z_k) \in \mathbb{R}^k \tag{2}$$

Table 4: An overview of the assumptions/limitations of MIAs found in the related literature.

Assumption/Limitation	Reference papers
MIAs in *white-box* setting only	[36, 48, 50]
MIAs reporting high success rates *only* on *overfitted* target ML models	[16, 25, 42, 43, 60, 64, 78, 79]
MIAs requiring certain (often private) information which might *not* be available to a potential attacker	[44, 48, 50, 59, 60, 64]
MIAs exploiting *only* a *small number* of vulnerable data records in the target model's training dataset	[35, 44, 67, 68]
MIAs having significantly *high overhead* on querying the target model for synthesizing training data records	[16, 64]
MIAs having significantly *high overhead* on querying the target model for conducting the attack	[8, 38]
MIAs *requiring* the use of shadow models that mimic the target model's behaviour	[5, 42–44, 60, 64, 78]
MIAs reporting high attack success rates *only* on *low-complexity* target ML models and datasets	[44, 64]

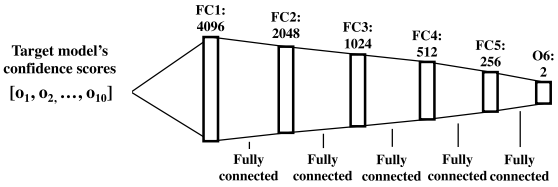


Figure 8: The adversary 2's fully-connected MLP. All the hidden layers utilize ReLU (Eq. (1)) as their activation function whereas the output layer uses Softmax (Eq. (2)). Finally, we use the SGD optimizer, a batch size of 64 samples, and the binary cross-entropy as the loss function.

$$\begin{aligned} Mean\ Squared\ Error\ (MSE) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \\ L1\ Loss &= \sum_{i=1}^{n}(|y_i - y_i'|) \\ L2\ Loss &= \sum_{i=1}^{n}(y_i - y_i')^2 \end{aligned} \tag{3}$$

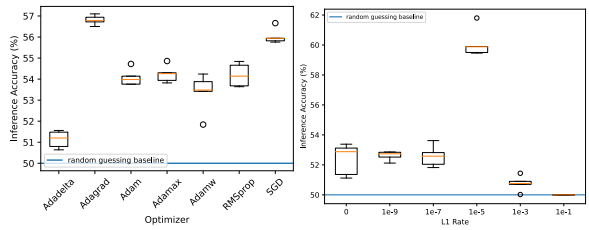
where y_i is the predicted output and y_i' is the target output of neuron i.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 + \boxed{\lambda \sum_{j=1}^{k}|w_j|} \tag{4}$$

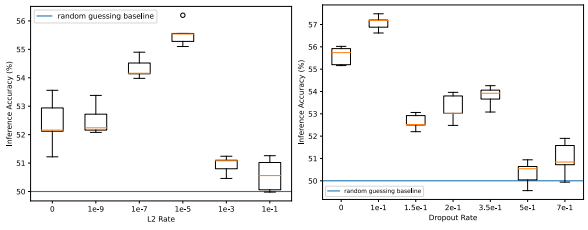
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 + \boxed{\lambda \sum_{j=1}^{k}w_j^2} \tag{5}$$

A.2 ML Generalization Mechanisms Evaluation

Figures 9 & 10 show the MIA success rates achieved on ResNet-34 trained on CIFAR-100 and SVHN, respectively, using different optimization and regularization mechanisms. In addition, Figures 11-13 show the MIA success rates achieved on MLP-1 trained on Adult, Surgical and IMDB, respectively, using different optimization and regularization mechanisms. As shown, the average MIA success rates remain close to the random guessing baseline (50%) for each tested setting. The target model's training/testing accuracies, when using each generalization technique, are shown in Tables 5-9 for CIFAR-100, SVHN, Adult, Surgical and IMDB, respectively.

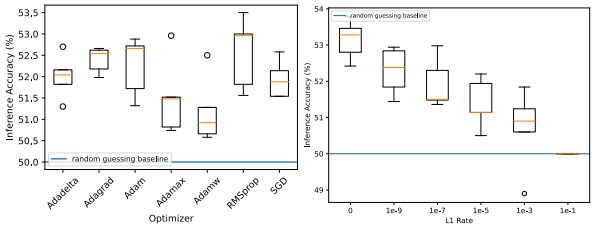


(a) Optimizer (b) L1 Rate

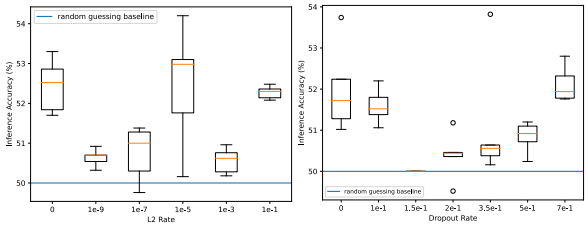


(c) L2 Rate (d) Dropout

Figure 9: MIA success rates on ResNet-34 trained on CIFAR-100 using various optimization and regularization methods.

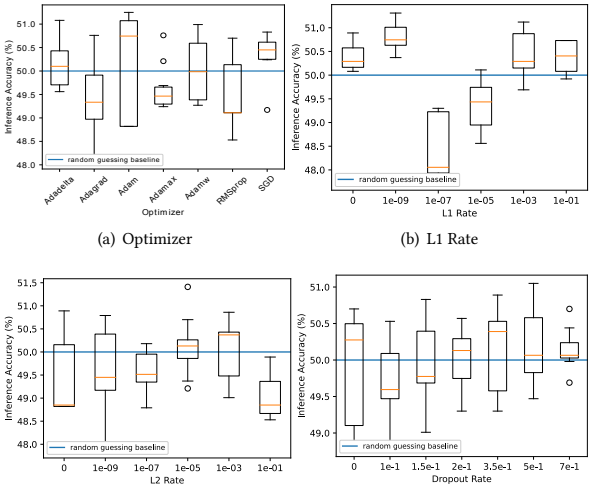


(a) Optimizer (b) L1 Rate



(c) L2 Rate (d) Dropout

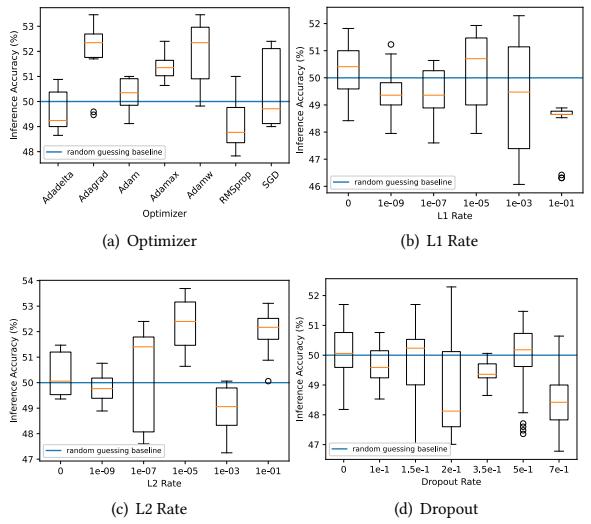
Figure 10: MIA success rates on ResNet-34 trained on SVHN using various optimization and regularization methods.



(a) Optimizer (b) L1 Rate

(c) L2 Rate (d) Dropout

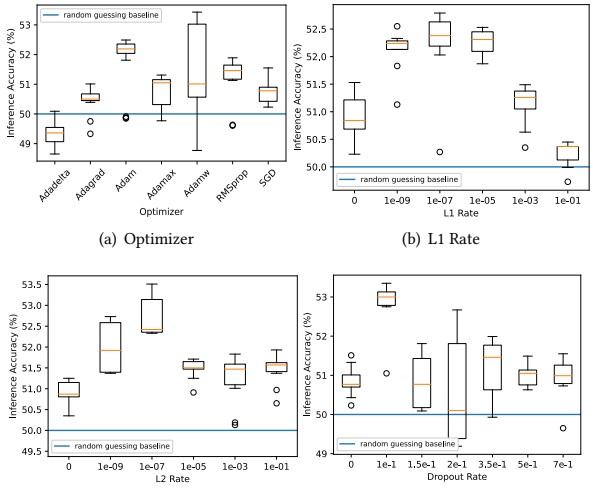
Figure 11: MIA success rates on MLP-1 trained on Adult using various optimization and regularization methods.



(a) Optimizer (b) L1 Rate

(c) L2 Rate (d) Dropout

Figure 12: MIA success rates on MLP-1 trained on Surgical using various optimization and regularization methods.



(a) Optimizer (b) L1 Rate

(c) L2 Rate (d) Dropout

Figure 13: MIA success rates on MLP-1 trained on IMDB using various optimization and regularization methods.

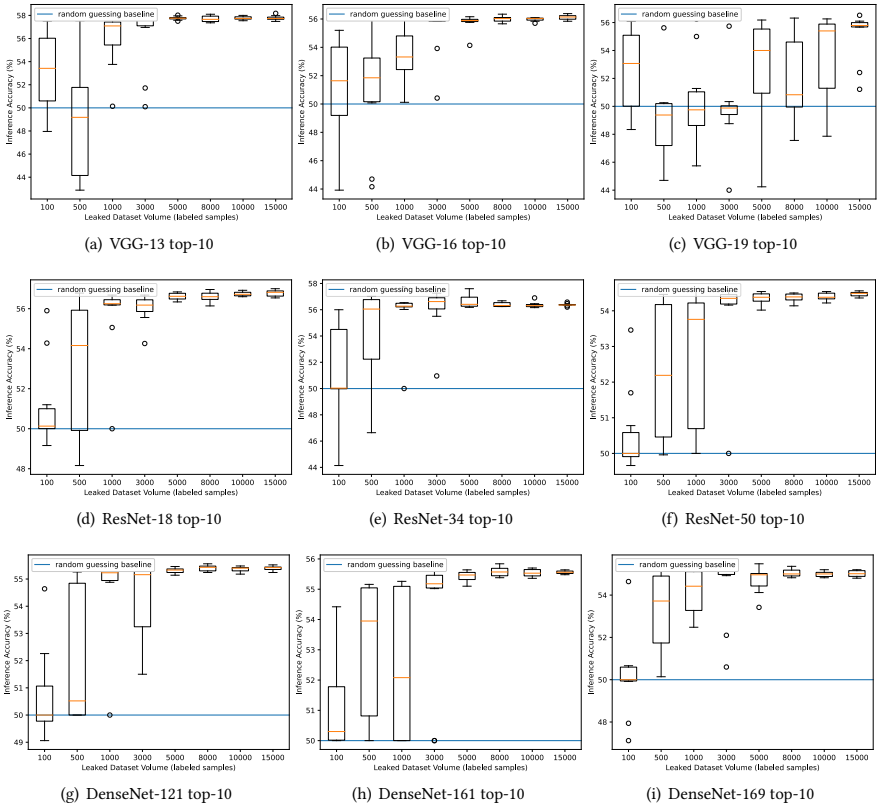
Table 5: ResNet-34 training/testing accuracies on CIFAR-100 for each optimization and regularization method used.

	Selection	Training Acc.	Testing Acc.
Optimizer	**SGD [32]**	53.13%	51.50%
	RMSprop [73]	60.22%	57.23%
	Adam [33]	58.15%	57.46%
	Adamax [33]	61.92%	59.50%
	Adadelta [81]	50.63%	43.26%
	Adagrad [12]	54.82%	49.26%
	AdamW [45]	59.93%	57.87%
***l*1 Rate**	**0**	55.88%	54.59%
	1e-9	69.11%	67.09%
	1e-7	69.21%	66.67%
	1e-5	70.30%	68.60%
	1e-3	38.66%	37.58%
	1e-1	14.56%	10.11%
***l*2 Rate**	**0**	58.81%	55.74%
	1e-9	57.58%	54.69%
	1e-7	62.35%	58.15%
	1e-5	66.69%	62.07%
	1e-3	63.32%	59.32%
	1e-1	17.45%	13.02%
Dropout Rate	**0**	59.99%	58.48%
	1e-1	61.25%	60.01%
	1.5e-1	58.22%	56.16%
	2e-1	59.32%	56.01%
	3.5e-1	53.51%	51.81%
	5e-1	43.96%	38.24%
	7e-1	34.48%	28.76%

A.3 Adversary 2 (Black-box) MIA Success Rates on Different ANNs and Datasets

This section shows adversary 2's MIA success rates achieved on different ANNs (with various depths) trained on datasets from various domains (CIFAR-10/100, SVHN, Adult, Surgical and IMDB), using top-k confidence scores + ground-truth membership status projection function. We omit the respective results for adversary 1 since they are similar to those observed from the (stronger) adversary 2.

As shown in Figures 14-19, all MIA success rates remain close to the random guessing baseline when targeting well-generalizable models. The collected results suggest that well-generalizable models are robust against black-box adversaries with strong background information (i.e., adversaries with access to the ground-truth membership status of a large portion of the target model's training set). Thus, Hui et al.'s observation (i.e., MIA success rate is analogous to the number of labeled instances available to the adversary) holds only for overfitted models and not for well-generalizable ones.



(a) VGG-13 top-10 (b) VGG-16 top-10 (c) VGG-19 top-10

(d) ResNet-18 top-10 (e) ResNet-34 top-10 (f) ResNet-50 top-10

(g) DenseNet-121 top-10 (h) DenseNet-161 top-10 (i) DenseNet-169 top-10

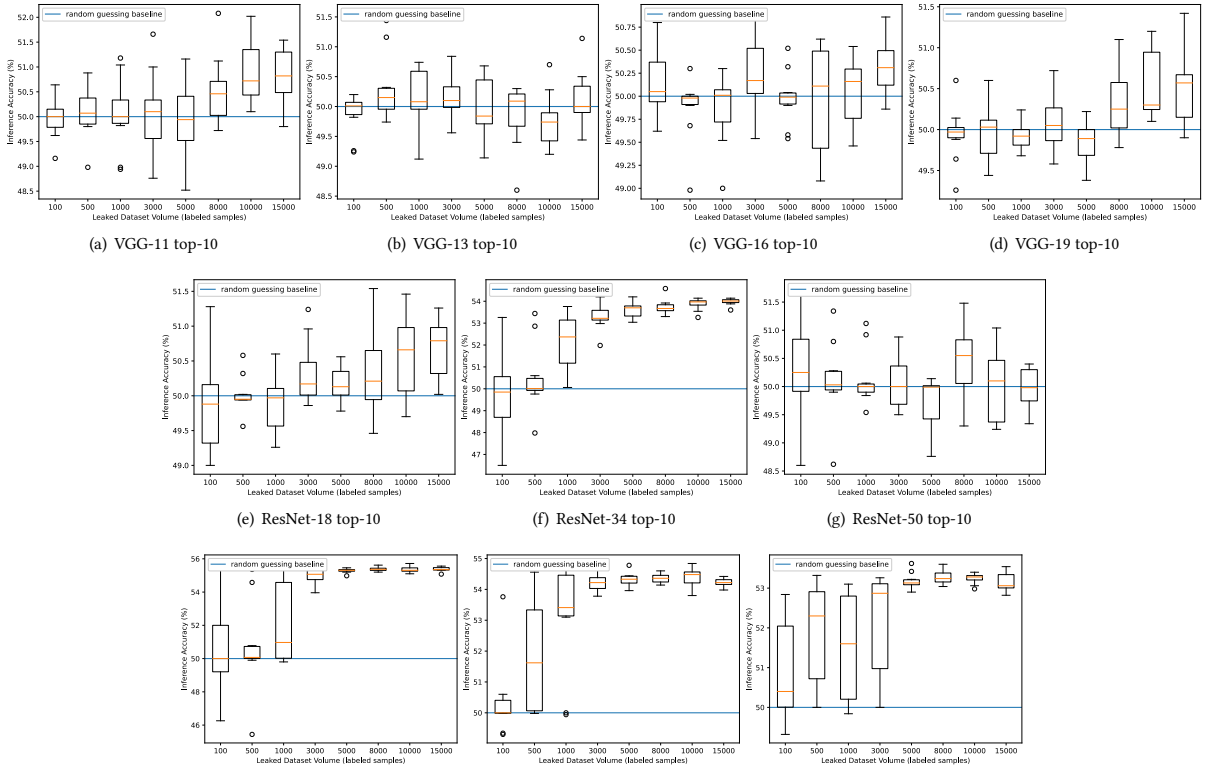
Figure 14: Adversary 2's MIA success rates achieved on VGG, ResNet and DenseNet (with various depths) trained on CIFAR-10. Since the collected results for $k = 3, 5, 7$ are similar to those observed for $k = 10$ we omit including the respective graphs.

Table 6: ResNet-34 training/testing accuracies on SVHN for each optimization and regularization method used.

	Selection	Training Acc.	Testing Acc.
Optimizer	**SGD [32]**	96.22%	95.07%
	RMSprop [73]	92.87%	90.24%
	Adam [33]	93.78%	93.21%
	Adamax [33]	95.88%	95.70%
	Adadelta [81]	92.45%	90.75%
	Adagrad [12]	94.76%	93.98%
	AdamW [45]	96.45%	96.32%
$l1$ Rate	**0**	96.01%	95.75%
	1e-9	96.89 %	95.50%
	1e-7	97.21%	96.04%
	1e-5	96.66%	95.75%
	1e-3	93.37%	91.13%
	1e-1	23.11%	19.58%
$l2$ Rate	**0**	96.42%	95.67%
	1e-9	96.85%	96.21%
	1e-7	96.99%	96.32%
	1e-5	96.45%	96.03%
	1e-3	95.87%	94.71%
	1e-1	80.23%	75.56%
Dropout Rate	**0**	97.88%	96.60%
	1e-1	97.33%	96.32%
	1.5e-1	97.45%	96.35%
	2e-1	97.02%	96.43%
	3.5e-1	97.11%	96.33%
	5e-1	96.25%	95.21%
	7e-1	91.48%	88.72%

Table 7: MLP-1 training/testing accuracies on Adult for each optimization and regularization method used.

	Selection	Training Acc.	Testing Acc.
Optimizer	**SGD [32]**	85.34%	85.30%
	RMSprop [73]	85.90%	85.47%
	Adam [33]	85.99%	84.81%
	Adamax [33]	85.74%	84.82%
	Adadelta [81]	56.03%	55.60%
	Adagrad [12]	80.52%	80.03%
	AdamW [45]	85.69%	85.24%
$l1$ Rate	**0**	85.34%	85.30%
	1e-9	85.62%	84.86%
	1e-7	86.20%	84.85%
	1e-5	86.04%	85.12%
	1e-3	85.59%	84.46%
	1e-1	75.34%	75.18%
$l2$ Rate	**0**	85.34%	85.30%
	1e-9	85.78%	85.11%
	1e-7	85.99%	85.24%
	1e-5	85.84%	85.55%
	1e-3	85.20%	85.05%
	1e-1	75.69%	75.23%
Dropout Rate	**0**	85.34%	85.30%
	1e-1	85.32%	85.19%
	1.5e-1	85.34%	84.89%
	2e-1	85.28%	85.10%
	3.5e-1	84.89%	84.58%
	5e-1	84.92%	83.54%
	7e-1	84.21%	82.60%



(a) VGG-11 top-10 (b) VGG-13 top-10 (c) VGG-16 top-10 (d) VGG-19 top-10

(e) ResNet-18 top-10 (f) ResNet-34 top-10 (g) ResNet-50 top-10

(h) DenseNet-121 top-10 (i) DenseNet-161 top-10 (j) DenseNet-169 top-10

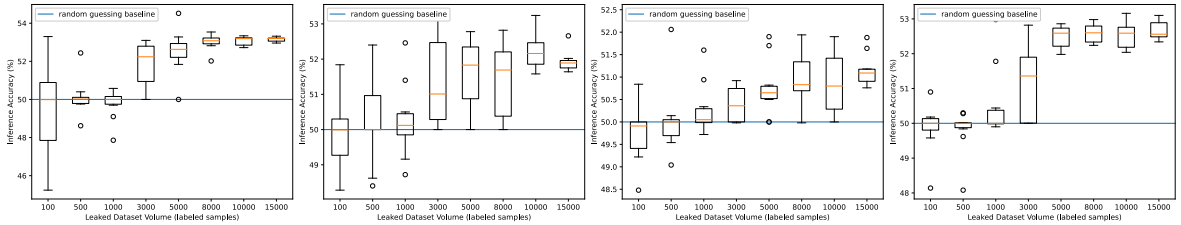
Figure 15: Adversary 2's MIA success rates achieved on VGG, ResNet and DenseNet (with various depths) trained on CIFAR-100. Since the collected results for $k = 3, 5, 7$ are similar to those observed for $k = 10$ we omit including the respective graphs.

Table 8: MLP-1 training/testing accuracies on Surgical for each optimization and regularization method used.

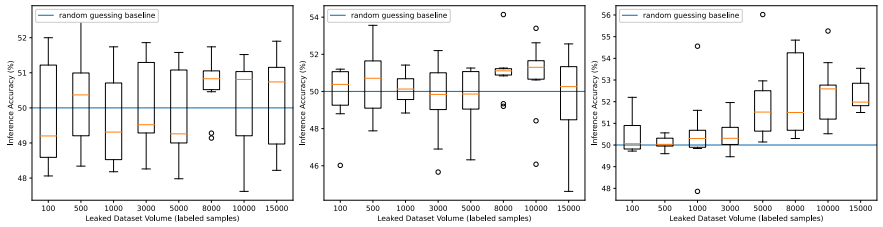
	Selection	Training Acc.	Testing Acc.
Optimizer	**SGD [32]**	80.10%	79.63%
	RMSprop [73]	80.53%	79.37%
	Adam [33]	80.91%	80.56%
	Adamax [33]	80.16%	78.94%
	Adadelta [81]	67.43%	66.42%
	Adagrad [12]	77.11%	76.07%
	AdamW [45]	80.73%	79.80%
***l1* Rate**	**0**	80.10%	79.63%
	1e-9	80.26%	79.29%
	1e-7	80.73%	79.29%
	1e-5	80.87%	79.12%
	1e-3	80.17%	80.05%
	1e-1	75.70%	74.48%
***l2* Rate**	**0**	80.10%	79.63%
	1e-9	80.79%	78.56%
	1e-7	80.61%	79.95%
	1e-5	80.18%	80.11%
	1e-3	80.91%	80.27%
	1e-1	74.70%	74.25%
Dropout Rate	**0**	80.10%	79.63%
	1e-1	80.41%	79.80%
	1.5e-1	80.06%	79.93%
	2e-1	79.92%	79.54%
	3.5e-1	80.40%	79.28%
	5e-1	80.05%	78.67%
	7e-1	79.88%	78.04%

Table 9: MLP-1 training/testing accuracies on IMDB for each optimization and regularization method used.

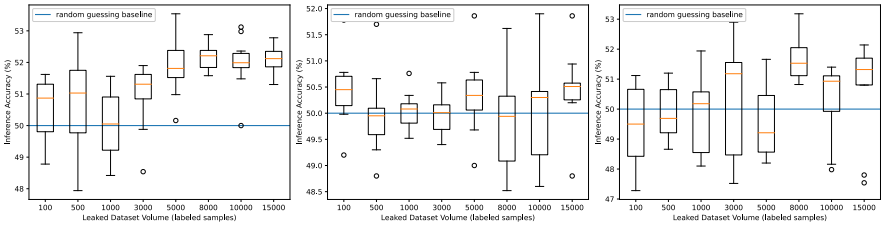
	Selection	Training Acc.	Testing Acc.
Optimizer	**SGD [32]**	92.17%	89.13%
	RMSprop [73]	91.07%	87.91%
	Adam [33]	92.27%	88.95%
	Adamax [33]	91.39%	89.49%
	Adadelta [81]	76.34%	75.97%
	Adagrad [12]	83.51%	83.35%
	AdamW [45]	90.57%	89.49%
***l1* Rate**	**0**	92.17%	89.13%
	1e-9	92.37%	89.16%
	1e-7	92.38%	88.87%
	1e-5	92.19%	89.20%
	1e-3	92.48%	88.98%
	1e-1	89.44%	87.84%
***l2* Rate**	**0**	92.17%	89.13%
	1e-9	92.39%	88.91%
	1e-7	92.41%	88.81%
	1e-5	92.23%	89.02%
	1e-3	92.31%	89.34%
	1e-1	91.17%	89.96%
Dropout Rate	**0**	92.17%	89.13%
	1e-1	92.18%	89.28%
	1.5e-1	91.86%	89.74%
	2e-1	91.96%	89.38%
	3.5e-1	91.47%	89.46%
	5e-1	90.09%	89.55%
	7e-1	87.02%	85.79%



(a) VGG-11 top-10 (b) VGG-13 top-10 (c) VGG-16 top-10 (d) VGG-19 top-10

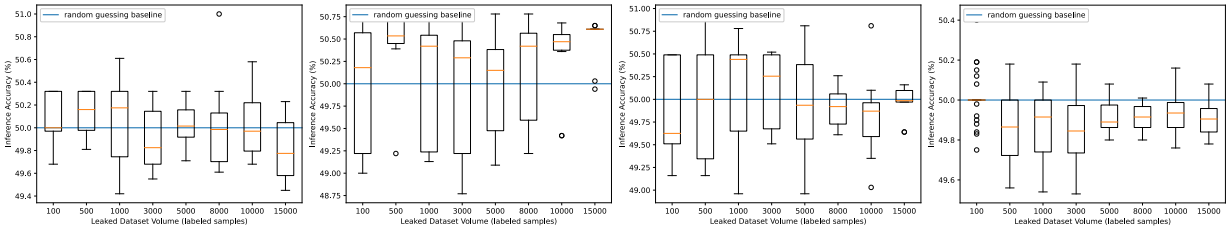


(e) ResNet-18 top-10 (f) ResNet-34 top-10 (g) ResNet-50 top-10



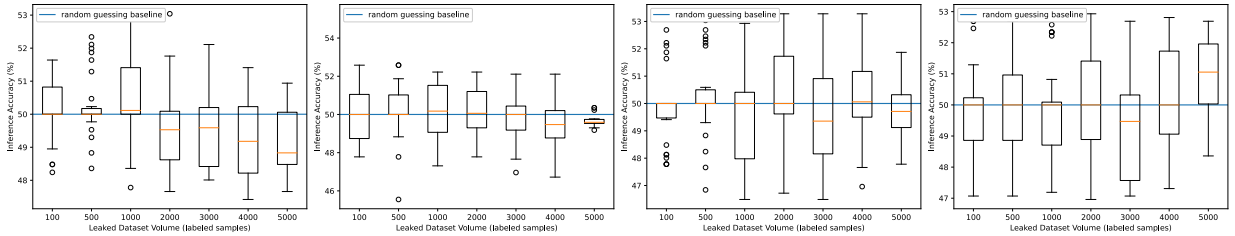
(h) DenseNet-121 top-10 (i) DenseNet-161 top-10 (j) DenseNet-169 top-10

Figure 16: Adversary 2's MIA success rates achieved on VGG, ResNet and DenseNet (with various depths) trained on SVHN. Since the collected results for $k = 3, 5, 7$ are similar to those observed for $k = 10$ we omit including the respective graphs.



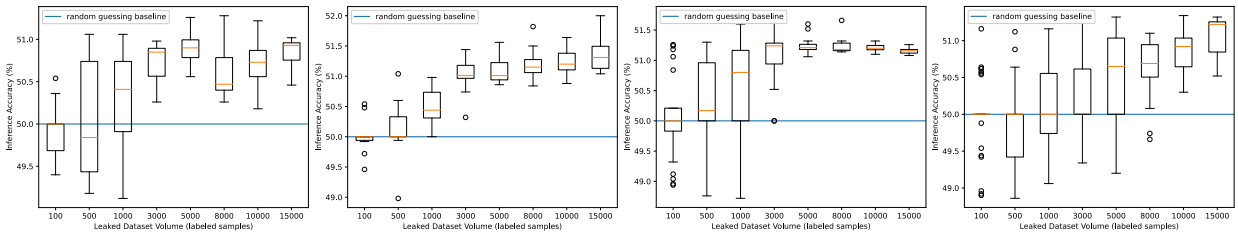
(a) MLP-1 top-2 (b) MLP-2 top-2 (c) MLP-3 top-2 (d) MLP-10 top-2

Figure 17: Adversary 2's MIA success rates achieved on MLP (with various depths, each layer having 10 neurons) trained on Adult. Since this dataset represents a binary classification problem $k = 2$.



(a) MLP-1 top-2 (b) MLP-2 top-2 (c) MLP-3 top-2 (d) MLP-10 top-2

Figure 18: Adversary 2's MIA success rates achieved on MLP (with various depths, each layer having 10 neurons) trained on Surgical. Since this dataset represents a binary classification problem $k = 2$.



(a) MLP-1 top-2 (b) MLP-2 top-2 (c) MLP-3 top-2 (d) MLP-10 top-2

Figure 19: Adversary 2's MIA success rates achieved on MLP (with various depths, each layer having 10 neurons) trained on IMDB. Since this dataset represents a binary classification problem $k = 2$.