To Trust or Not To Trust Prediction Scores for Membership Inference Attacks

Dominik Hintersdorf [*1] , **Lukas Struppek** [*1] , **Kristian Kersting** [1,2,3]

[1]Department of Computer Science, Technical University of Darmstadt, Germany
[2]Centre for Cognitive Science, Technical University of Darmstadt, Germany
[3]Hessian Center for AI (hessian.AI), Germany
{dominik.hintersdorf, lukas.struppek, kersting}@cs.tu-darmstadt.com,

Abstract

Membership inference attacks (MIAs) aim to determine whether a specific sample was used to train a predictive model. Knowing this may indeed lead to a privacy breach. Most MIAs, however, make use of the model's prediction scores—the probability of each output given some input—following the intuition that the trained model tends to behave differently on its training data. We argue that this is a fallacy for many modern deep network architectures. Consequently, MIAs will miserably fail since overconfidence leads to high false-positive rates not only on known domains but also on out-of-distribution data and implicitly acts as a defense against MIAs. Specifically, using generative adversarial networks, we are able to produce a potentially infinite number of samples falsely classified as part of the training data. In other words, the threat of MIAs is overestimated, and less information is leaked than previously assumed. Moreover, there is actually a trade-off between the overconfidence of models and their susceptibility to MIAs: the more classifiers know when they do not know, making low confidence predictions, the more they reveal the training data.

1 Introduction

Deep learning models achieve state-of-the-art performances in various tasks such as computer vision, language modeling, and healthcare. However, large datasets are needed to train these models. Collecting and, in particular, cleaning and labeling data is expensive. Hence, users may look for alternative data sources, which may not always be legal ones. To detect data abuse, it would be desirable to prove whether a model was trained on leaked or unauthorized retrieved data. However, to prove that a specific data point was part of the training set is difficult since neural networks do not store plain training data like lazy learners. Instead, the learned knowledge is encoded into the network's weights.[1]

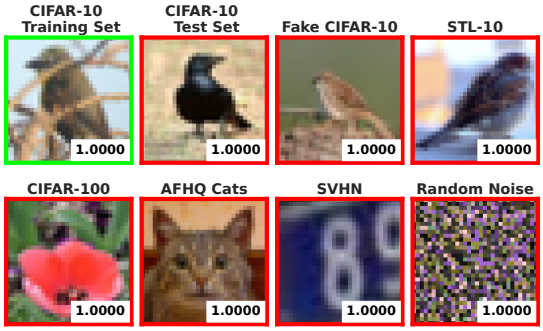


Figure 1: False-positive membership inference attacks (red frames) against a ResNet-18 and their assigned maximum prediction scores.

One way to distinguish between unseen data and data points used for training the neural networks is through membership inference attacks (MIAs). They attempt to identify training samples in a large set of possible inputs. Besides malicious intentions, MIAs might be used to prove illegal data abuse in deep learning settings. To use membership inference results as evidence in court, high accuracy and robustness to different data types and network architectures is required.

Previous works on MIAs, see e.g., Shokri et al. [2017] and Salem et al. [2019], state strong attack results in distinguishing between training and test data, and give the impression that MIAs have a strong impact on a model's privacy. However, the evaluation of MIAs reported in the literature is usually done with limited data in a cross-validation setting, i.e., on samples from the exact same data distribution, not considering other distributions with possibly similar image contents.

We argue that MIAs, in particular attacks based on a model's prediction scores, are not robust and not very meaningful in realistic settings, due to their high false-positive rates, also criticized by Rezaei et al. [2021]. We take, however, a broader view and do not restrict evaluation on the target model's exact training distribution. In a specific domain, there is a possibly infinite number of samples and hence the number of false positives can be increased arbitrarily. This leads to reduced informative value and low reliability of the attacks under realistic conditions. Fig. 1 shows samples from various datasets for which all three MIAs studied in this paper make false-positive predictions, even if the inputs are nothing similar to the training data or do not contain any meaningful information at all. We practically demonstrate the theo-

[*]Equal contribution.
[1]Extended paper available at https://arxiv.org/abs/2111.09076.

retically unlimited number of false-positive member classifications by using a GAN to generate images following the training distribution.

Our argumentation is based on the already known overconfidence of modern deep neural architectures [Nguyen *et al.*, 2015; Hendrycks and Gimpel, 2017; Guo *et al.*, 2017; Leibig *et al.*, 2019]. However, overconfidence has consistently been ignored in the MIA literature, even though MIA findings are already having an impact on regulatory and other legal measures. Our experimental results indicate that mitigating the overconfidence of neural networks using calibration techniques increases privacy leakage.

We argue that previous works performed misleading attack evaluations and overestimated the actual attack effectiveness by using only data from the target model's exact training distribution. Actually, there might not exist any meaningful MIA at all since the attacks will always produce a high number of false positives due to the overconfidence of neural networks.

To summarize, we make the following contributions:

1. We demonstrate that the effectiveness of MIAs has been systematically overestimated by ignoring the fact that most neural networks are inherently overconfident and, therefore, produce high false-positive rates.
2. We show that overconfidence acts as a natural defense against MIAs.
3. We reveal that a trade-off exists between keeping models secure against MIAs and mitigating overconfidence.

We proceed as follows. We start off by reviewing MIAs and how overconfidence of neural networks can be mitigated. Afterward, we introduce the theoretical background and our experimental setup. Before discussing and concluding our work, we present our experimental results.

2 Membership Inference Attacks

Membership inference attacks (MIAs) on neural networks were first introduced by Shokri et al. [2017]. In a general MIA setting, as usually assumed in the literature, an adversary is given an input x following distribution D and a target model M_{target} which was trained on a training set $S^{target}_{train} \sim D^n$ with size n. The adversary is then facing the problem to identify whether a given $x \sim D$ was part of the training set S^{target}_{train}. To predict the membership of x, the adversary creates an inference model h. In score-based MIAs, the input to h is the prediction score vector produced by M_{target} on sample x. Since MIAs are binary classification problems, precision, recall, false-positive rate (FPR), and area under the receiver operating characteristic (AUROC) are used as attack evaluation metrics in our experiments.

All MIAs exploit a difference in the behavior of M_{target} on seen and unseen data. Most attacks in the literature follow Shokri et al. [2017] and train so-called shadow models M_{shadow} on a disjoint dataset S^{shadow}_{train} drawn from the same distribution D as S^{target}_{train}. M_{shadow} is used to mimic the behavior of M_{target} and adjust parameters of h, such as threshold values or model weights. Note that the membership status for inputs to M_{shadow} are known to the adversary. Fig. 2 visualizes the attack preparation process.

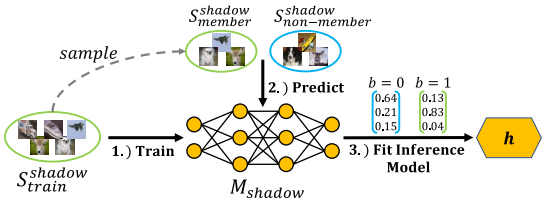


Figure 2: Membership inference preparation process.

In recent years, various MIAs have been proposed. Shokri et al. [2017] trained multiple shadow models and queried each of the shadow models with its training data (members), as well as unseen data (non-members) to retrieve the prediction scores of the shadow models. Multiple binary classifiers were then trained for each class label to predict the membership status. Salem et al. [2019] also used prediction scores and trained a single class-agnostic neural network to infer membership. In contrast to Shokri et al. [2017], their approach relies on a single shadow model. The input of h consists of the k highest prediction scores in descending order.

Instead of focusing solely on the scores, Yeom et al. [2018] took advantage of the fact that the loss of a model is lower on members than on non-members and fit a threshold to the loss values. More recent approaches [Choquette-Choo *et al.*, 2021; Li and Zhang, 2021] focused on label-only attacks where only the predicted label for a known input is observed.

3 Overconfidence of Neural Networks

Neural networks usually output prediction scores, e.g., by applying a softmax function. To take model uncertainty into account, it is usually desired that the prediction scores represent the probability of a correct prediction, which is usually not the case. This problem is generally referred to as model calibration. Guo et al. [2017] demonstrated that modern networks tend to be overconfident in their predictions. Hein et al. [2019] have further proven that ReLU networks are overconfident even on samples far away from the training data.

Existing approaches to mitigate overconfidence can be grouped into two categories: post-processing methods applied on top of trained models and regularization methods modifying the training process. As a post-processing method, Guo et al. [2017] proposed temperature scaling using a single temperature parameter T for scaling down the pre-softmax logits for all classes. The larger T is, the more the resulting scores approach a uniform distribution. Kristiadi et al. [2020] further proposed to approximate a model's final layer with a Laplace approximation. Müller et al. [2019] demonstrated that label smoothing regularization Szegedy et al. [2016] not only improves the generalization of a model but also implicitly leads to better model calibration. The calibration of a model can be measured by the expected calibration error (ECE) [Naeini *et al.*, 2015] and the overconfidence error (OE) [Thulasidasan *et al.*, 2019]. Both metrics compute a weighted average over the absolute difference between test accuracy and prediction scores, while ECE penalizes the calibration gap and OE penalizes overconfidence.

4 Do Not Trust Prediction Scores for MIAs

In this section, we will show that predictions scores for MIAs cannot be trusted because score-based MIAs make membership decisions based mainly on the maximum prediction score. As a first step, we mathematically motivate our argumentation and then verify our claims empirically.

Formally, a neural network $f(x)$ using ReLU activations decomposes the unrestricted input space $\mathbb{R}^m$ into a finite set of polytopes (linear regions). We can then interpret $f(x)$ as a piecewise affine function that is affine in any polytope [Arora *et al.*, 2018]. Due to the limited number of polytopes, the outer polytopes extend to infinity which allows to arbitrarily increase the prediction scores through scaling inputs by a large constant δ [Hein *et al.*, 2019]. We now further develop these findings from an MIA point of view and state the following theorem:

Theorem 1. *Given a (leaky) ReLU-classifier, we can force almost any non-member input to be classified as a member by score-based MIAs, simply by scaling it by a large constant.*

Proof. Let $f : \mathbb{R}^m \rightarrow \mathbb{R}^d$ be a piecewise affine (leaky) ReLU-classifier. We define a score-based MIA inference model $h : \mathbb{R}^d \rightarrow \{0, 1\}$ with 1 indicating a classification as a member. For almost any input $x \in \mathbb{R}^m$ and a sufficiently small $\epsilon > 0$ if $\max_{i=1,\dots,d} \ f(x)_i \geq 1 - \epsilon$, it follows that $h(f(x)) = 1$. Since $lim_{\delta\rightarrow\infty} \max_{i=1,\dots,d} f(\delta x)_i = 1$, then $lim_{\delta\rightarrow\infty} h(f(\delta x))= 1$ already holds. □

By scaling the whole non-member dataset, one can force the FPR to be close to 100%. Indeed, the theorem holds only for (leaky) ReLU-networks and unbounded inputs. However, since uncalibrated neural networks assign high prediction scores to a wide range of different inputs, the number of false-positive predictions is also large for unscaled inputs from known and unknown domains. Next, we empirically show that one cannot trust predictions scores for MIAs in more general settings without input scaling required and using other activation functions.

4.1 Experimental Protocol

We make our source code publicly available[2].

Threat Model. As in most MIA literature [Salem *et al.*, 2019; Yeom *et al.*, 2018; Song and Mittal, 2021], we followed the MIA setting of Shokri et al. [2017], and like Salem et al. [2019] only trained a single shadow model for each attack. As in previous work, we also simulate a worst-case scenario, i.e., the adversary knows the exact architecture and training procedure of the target model. Therefore, a strong shadow model can be trained, following the procedure depicted in Fig. 2. In our score-based MIA scenario, the adversary only has access to the target model's prediction scores.

Datasets. We evaluated the attacks on models trained on the CIFAR-10 [Krizhevsky, 2009] and Stanford Dogs [Khosla *et al.*, 2011] datasets.

[2]Available at https://github.com/ml-research/To-Trust-or-Not-To-Trust-Prediction-Scores-for-Membership-Inference-Attacks

	SalemCNN	**ResNet-18**	**EfficientNetB0**
Train Accuracy	100.00%	100.00%	99.03%
Test Accuracy	59.04%	69.38%	71.06%
Entropy Pre	65.51%	67.35%	61.36%
Entropy Rec	88.52%	92.32%	79.96%
Entropy FPR	46.60%	44.76%	50.36%
Entropy AUROC	70.94%	76.50%	66.57%
Max. Score Pre	65.34%	67.35%	61.43%
Max. Score Rec	87.48%	92.32%	79.64%
Max. Score FPR	46.40%	44.76%	50.00%
Max. Score AUROC	72.03%	77.50%	66.58%
Top-3 Scores Pre	62.48%	63.84%	60.74%
Top-3 Scores Rec	100.00%	98.04%	82.60%
Top-3 Scores FPR	60.04%	55.52%	53.40%
Top-3 Scores AUROC	71.57%	77.14%	66.61%

Table 1: Training and attack metrics for the target models trained on CIFAR-10. We measure the attacks' precision (Pre), recall (Rec), FPR and AUROC on equally-sized member and non-member subsets from CIFAR-10.

We created two disjoint training datasets for the target and shadow models, each containing 12,500 (CIFAR-10) and 8,232 (Stanford Dogs) samples. We then randomly drew 2,500 and 2,058 samples, respectively, from the training and test sets to create the member and non-member datasets.

We used various datasets to demonstrate the susceptibility of prediction score-based MIAs to high scores on samples from neighboring distributions and samples further away from the training data—a kind of out-of-distribution (OOD) setting. We used STL-10 [Coates *et al.*, 2011], CIFAR-100 [Krizhevsky, 2009], SVHN [Netzer *et al.*, 2011], and Animal Faces-HQ (AFHQ) [Choi *et al.*, 2020] as datasets.

Additionally, we used pre-trained StyleGAN2 [Karras *et al.*, 2020] models to generate synthetic CIFAR-10 and dog images, referred to as Fake CIFAR-10 and Fake Dogs. To empirically verify our theorem and push our approach to the extreme, we created two additional datasets based on the respective test images by scaling pixel values by factor 255 and randomly permuting the images' pixels to create random noise samples. In the following, we refer to these two datasets as Permuted and Scaled.

Neural Network Architectures. On CIFAR-10, we trained a ResNet-18 [He *et al.*, 2016], an EfficientNetB0 [Tan and Le, 2019] and a simple convolutional neural network following Salem et al. [2019], referred to as SalemCNN. For the Stanford Dogs dataset, we used a larger ResNet-50 architecture pre-trained on ImageNet. ResNets and SalemCNN are ReLU networks and can be interpreted as piecewise linear functions [Arora *et al.*, 2018]. EfficientNetB0 uses Swish activation functions [Ramachandran *et al.*, 2018], which are not piecewise linear and, therefore, our theorem does not hold. Nevertheless, we demonstrate that also non-ReLU networks suffer from overconfidence, leading to weak MIAs.

Prediction Score-Based Attacks. We base our analysis on three different MIAs [Salem *et al.*, 2019] exploiting the top-3 values of the prediction score vector, the maximum value, and the entropy. For the top-3 prediction score attack, we trained a small neural network with a single hidden layer as an infer-

ence model. It uses the three highest scores of M_{target} in descending order as inputs. The maximum prediction score attack relies only on the highest score, while the entropy attack computes the entropy on the whole prediction score vector. An input sample is classified as a member, if the maximum value is higher or if the entropy is lower than a threshold. We fit all attack models on the shadow models' outputs, with the thresholds chosen to maximize the true-positive rate while minimizing the FPR.

4.2 Experimental Results

We investigate the following questions: **(Q1)** How robust are prediction score-based MIAs? **(Q2)** Does overconfidence negatively affect MIAs? **(Q3)** How does calibrating neural networks influence the success of MIAs? **(Q4)** Are defenses contrary to calibration?

(Q1) MIAs Are Not Robust. Tab. 1 summarize the test accuracy and attack metrics of the CIFAR-10 target models. The different attacks performed quite similarly while the recall is always significantly higher than the precision, indicating the problem of many false-positive predictions. A similar picture emerges when looking at the results of the Standard Stanford Dog model, stated in Tab. 2.

To examine the robustness of the attacks, we used the remaining datasets as non-member inputs and measured the FPRs. Figs. 3b and 3d (transparent bars), show the FPR of the attacks against the ResNet CIFAR-10 models, and Figs. 3a and 3c do the same for the Stanford Dogs models.

The results demonstrate that the attacks not only tend to falsely classify samples from the test data as members but also samples from other distributions. For example, the attacks against CIFAR-10 misclassified more than a third of the STL-10 samples, which are similar in content and style, as members. The same holds for AFHQ Dogs samples as input for the Stanford Dogs model. The results on the remaining datasets, especially on the scaled samples, empirically confirm our theorem and demonstrate that neural networks are not able to recognize when they are operating on unknown inputs, such as housing numbers, cats, or random noise, and therefore still produce high FPRs. Even on generated Fake samples following the training distribution, the FPR is comparably high and shows that there exists a potentially infinite number of false-positive samples that are not out-of-distribution. This behavior is not limited to ReLU networks. The FPR of the EfficientNetB0 on the datasets is quite similar to the FPR of the ResNet-18. This indicates that the problem of high FPR in MIAs is affecting modern deep architectures in general and underlines the fact that MIAs are not robust.

(Q2) High Prediction Scores Lower Privacy Risks. To shed light on the connection between overconfidence and high FPR of the MIAs, we analyzed the mean maximum prediction scores (MMPS) of the target models' predictions.

Tab. 3 shows the MMPS values measured on a standard ResNet-50 and underlines our assumption that all score-based MIAs against models trained with standard procedure mainly rely on the maximum score since there is a clear difference between the MMPS of false-positive and true-negative predictions. The results we have obtained for the CIFAR-10 models are similar to the Results on the ResNet-50.

		Calibration		Defenses	
ResNet-50	**Standard**	**LS**	**LA**	**Temp**	**L2**
Train Accuracy	98.48%	99.62%	98.48%	98.48%	74.05%
Test Accuracy	59.69%	64.65%	59.62%	59.69%	48.15%
ECE	**25.09%**	**↓5.80%**	5.63%	51.03%	11.86%
OE	**21.18%**	**↓0.32%**	3.59%	0.0%	7.83%
Entropy Pre	68.22%	76.33%	65.39%	59.45%	60.50%
Entropy Rec	84.50%	82.56%	87.03%	47.38%	50.68%
Entropy FPR	**39.36%**	**↓25.61%**	46.06%	32.31%	33.09%
Entropy AUROC	**78.22%**	**↑85.41%**	77.96%	**↓60.84%**	**↓61.40%**
Max. Score Pre	68.30%	77.32%	68.44%	63.96%	59.13%
Max. Score Rec	83.97%	81.83%	83.87%	65.55%	56.66%
Max. Score FPR	**38.97%**	**↓24.00%**	38.68%	36.93%	39.16%
Max. Score AUROC	**78.12%**	**↑85.63%**	78.15%	**↓69.80%**	**↓61.84%**
Top-3 Scores Pre	67.48%	76.36%	67.88%	68.48%	59.41%
Top-3 Scores Rec	85.81%	85.71%	84.60%	85.08%	55.39%
Top-3 Scores FPR	**41.35%**	**↓26.53%**	40.04%	39.16%	37.85%
Top-3 Scores AUROC	**78.29%**	**↑86.24%**	78.38%	79.60%	**↓61.86%**

Table 2: Training and attack metrics for ResNet-50 target models trained on Stanford Dogs. We compare the results for the standard model to models trained with label smoothing (LS) and Laplace approximation (LA) as calibration techniques and temperature scaling (Temp) and L2 regularization as defense techniques. Arrows indicate the differences compared to the standard model.

It seems that the non-maximum scores are not providing significant information on the membership status since the MMPS values of the false-positive predicted samples using the maximum score attack and the top-3 attack differ only slightly. Modifying the top-3 attack to use a larger part of the score vector for inferring membership of the samples did not significantly improve the membership inference either.

So on one side, neural networks are overconfident in their predictions, even on inputs without any known content. It prevents a reasonable interpretation regarding a model's probability of being correct in its predictions. During MIAs, on the other side, this behavior implicitly protects the training data since the information content of the prediction score is rather low. Consequently, there is a trade-off between a model's ability to react to unknown inputs and its privacy leakage. We explore this trade-off in Q3. We further argue that any adversarial example maximizing the target model's scores in an arbitrary class would also be classified as a member in almost all cases. So it is possible to hide members in a larger dataset of non-members that are altered by adversarial attacks to maximize the target model's scores.

(Q3) Mitigating Overconfidence Increases Privacy Risks. Ideally, neural networks are properly calibrated, and their prediction scores represent the probabilities of correct predictions. To calibrate the models and to reduce the overconfidence, we retrained the ResNet-18 and ResNet-50 models with label smoothing. We performed the same calibration method on both the target and the shadow models, which reflects a worst-case scenario, with an adversary knowing the exact calibration method and hyperparameters.

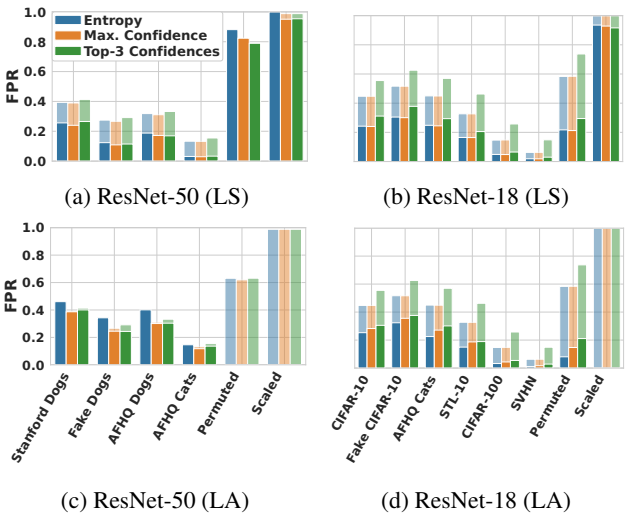
Label smoothing not only calibrates a model but may also improve its test accuracy, as shown in Tab. 2 for ResNet-50.

Both the expected calibration error (ECE) and overconfidence error (OE) dropped significantly, demonstrating a strong calibration effect when using label smoothing.

Previous works on MIAs suggested that minimizing the accuracy gap between the training and test accuracy on the

Dataset	Attack	FP MMPS	TN MMPS
Stanford Dogs	Entropy	0.9984	0.7565
	Max. Score	0.9985	0.7580
	Top-3 Scores	0.9979	0.7486
Fake Dogs	Entropy	0.9977	0.7700
	Max. Score	0.9979	0.7724
	Top-3 Scores	0.9971	0.7648
AFHQ Cats	Entropy	0.9972	0.7205
	Max. Score	0.9972	0.7208
	Top-3 Scores	0.9959	0.7137

Table 3: MMPS for false-positive (FP) and true-negative (TN) predictions of different attacks on the standard ResNet-50 model on selected datasets. A clear difference between false-positive and true-negative mean maximum prediction scores for all attacks can be seen. This indicates that all of the analyzed attacks heavily relied on the maximum prediction score.



(a) ResNet-50 (LS) (b) ResNet-18 (LS)

(c) ResNet-50 (LA) (d) ResNet-18 (LA)

Figure 3: False-positive rates (FPR) of MIAs against ResNet-18 and ResNet-50. The transparent bars represent the FPR of the standard models, whereas the solid bars represent the FPR of the models with the respective modification given in parentheses - label smoothing (LS) and Laplace approximation (LA). Both calibration methods reduce the FPR for almost all inputs.

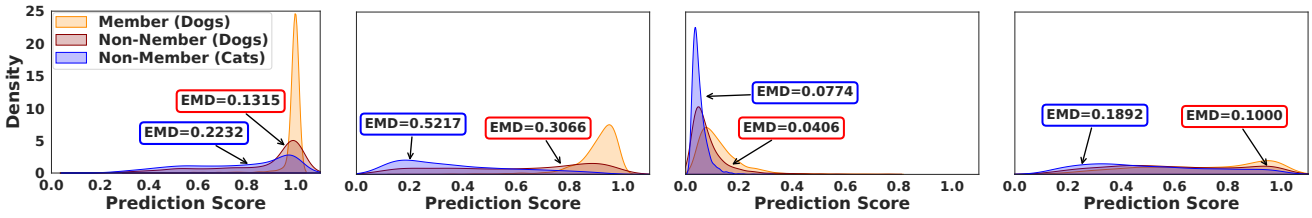
same architecture leads to weaker attacks and, therefore, to lower privacy risks. However, as demonstrated by the results summarized in Tab. 2, label smoothing improves the test accuracy and still yields higher attack precision values for all three attacks on both architectures. Figs. 3a and 3b further illustrate that label smoothing reduces the number of false-positive membership predictions. Whereas the FPR on the Permuted samples is drastically reduced for ResNet-18, the FPR of the ResNet-50 on the Permuted samples even increases when using label smoothing. We note that this effect does only occur in some training runs. In other cases, the FPR for Permuted data drops similar to the ResNet-18 results. On all datasets, the reductions in the FPR are comparable between the ResNet-18 and ResNet-50. The FPR also decreases for inputs similar to the training data. For comparison, we also apply a Laplace approximation (LA) on the weights of the final layers to mitigate overconfidence. As shown in Figs. 3c and 3d, LA is better suited to avoid high prediction scores on the Permuted and Scaled samples.

Our results demonstrate that if a model shows reduced prediction scores on unseen inputs, the samples of the training data are easier to identify. It reduces the protection induced by overconfident predictions (on unseen inputs) and increases vulnerability to MIAs. We applied a kernel density estimation (KDE) to visualize the distribution of the maximum prediction scores of the ResNet-50 target models on member and non-member data. Figs. 4a and 4b show the estimated density functions. Without label smoothing, all three distributions have their mode around prediction scores of 1.0. This leads to a large overlap of the distributions. Samples with prediction scores this high are most likely classified as false-positive members as the FP MMPS values in Tab. 3 suggest. We also state the earth mover's distance (EMD) in the KDE plots to quantify the distance between the member and non-member distributions. Label smoothing separates the three distributions clearly and doubles both EMD values. The label smoothing model tends to be less overconfident in its predictions on unknown input data, and hence the member samples are easier to separate from non-members. This increases the potential privacy leakage of MIAs.

As depicted in Fig. 5, we further used t-SNE [van der Maaten and Hinton, 2008] to plot the penultimate layer activations on samples from the same datasets as used for the KDE plots. Whereas the standard model in Fig. 5a shows an overlapping between the activations of the three datasets, label smoothing in Fig. 5b creates tighter clusters of dog samples and separates the OOD cat images more clearly.

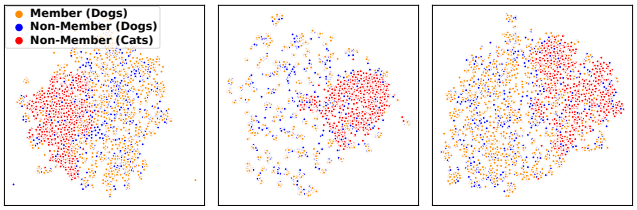
(Q4) A Trade-off Between Calibration and Defenses Exists. Whereas calibration tries to maximize the informative value of the prediction scores, many defenses against MIAs aim to reduce the informative value and to align the score distributions of members and non-members. In this section, we want to investigate whether it is possible to defend calibrated models or a trade-off between calibration and defenses against MIAs exists. Defenses reduce the generalization of a model in terms of its ability to distinguish between samples from known and unknown inputs and express meaningful scores. To test this, we first applied temperature scaling with $T = 10$ to the trained ResNet-50 standard model without calibration. Fig. 4c shows the estimated maximum prediction score distributions. The score vectors converge to a uniform distribution, and the distributions of the top scores are much more similar. This can be seen by the significantly lower EMD values. With an ECE of 51% using temperature scaling, the information content of the actual prediction score is greatly reduced, and the AUROC for the Entropy and Maximum Score attacks drop significantly, as shown in Tab. 2. On the top-3 score attack, temperature scaling has no effect. We suspect this is due to the added temperature term being a monotone transformation, not removing information encoded in the top-3 score patterns.

We also investigated L2 regularization as a stronger defense applied during training on our ResNet models. L2 regularization effectively reduces the vulnerability to MIAs.



(a) ResNet-50 (b) ResNet-50 (LS) (c) ResNet-50 (Temperature) (d) ResNet-50 (L2)

Figure 4: Kernel density estimation applying Gaussian kernels on the top prediction scores values of ResNet-50 target models. We use equally-sized member and non-member subsets of Stanford Dogs and AFHQ Cats. We further state the earth mover's distance (EMD) between each dataset and the member dataset. Label smoothing (LS) moves the non-member distributions further away, and consequently, the members become easier to separate. Temperature scaling and L2 regularization show an inverse effect and increase the overlapping.



(a) ResNet-50 (b) ResNet-50 (LS) (c) ResNet-50 (L2)

Figure 5: T-SNE visualization of the penultimate ResNet-50 layer activations on training samples (orange), test samples (blue), and OOD samples (red). Label smoothing (LS) creates much tighter clusters of training and OOD cat samples, which makes them easier to separate, whereas L2 regularization has a reverse effect.

For all attacks, both precision and recall drop significantly at the cost of reduced test accuracy, as Tab. 2 states. Moreover, the ECE and OE are significantly higher than for the model trained with label smoothing. The distribution of the highest prediction scores can be seen in Fig. 4d. Similar to temperature scaling, L2 regularization aligns the distributions of members and non-members but distributes the maximum scores more equally instead of pushing it towards a single value. Fig. 5c shows a similar effect of overlapping distributions in the penultimate layer activations, making it harder to separate members from non-members and OOD data.

As shown in our experiments, defenses are contrary to calibration. Our results indicate that a trade-off exists between defending models against MIAs and applying calibration to increase the model's informative value.

5 Discussion

In all our analyses, we followed the standard threat model for MIAs in the literature and assumed a strong adversary with full knowledge about the target model's architecture and training procedure and having access to data from the target's training distribution. Our experiments underline the known fact that modern neural networks are not inherently able to identify unseen and unknown inputs and cannot adapt their behavior in terms of reducing the prediction scores. However, we have shown that this is why the expressiveness of MIAs in realistic scenarios is greatly reduced, and the associated privacy risks are thus much lower than previously assumed. Loosening the attack scenario assumptions and providing the attacker with even less information during an attack, the effectiveness of MIAs will decrease even further.

One way to mitigate the problem of false-positive predictions on unseen data is to first try to identify and remove all OOD samples. This would indeed prevent some false-positive predictions caused by completely different data distributions. However, we demonstrated that the problem of high FPR also occurs on datasets similar to the training data. In this case, the adversary has no means to tell whether a given sample is in- or out-of-distribution if the images' contents are similar, which in turn makes it impossible for the attacker to filter out OOD samples. Even if this were possible, by generating synthetic images, we have shown that there is a potentially unlimited number of samples that follow the training distribution and still lead to false-positive MIA predictions, questioning the overall informative value of MIAs.

We only considered prediction score-based MIAs, but we expect our results to be similar for other kinds of attacks. Doing so provides an interesting avenue for future work. Also, future research should further investigate the trade-off between MIA defenses and calibration of machine learning models and how both aspects could be balanced. Furthermore, including techniques from open set recognition and OOD detection into MIAs might improve their effectiveness.

6 Conclusion

We have shown that MIAs produce high false-positive rates due to overconfident predictions of modern neural networks for in- and out-of-distribution data. In stark contrast to previous works stating strong attack results on standard neural networks, we demonstrate that MIAs are actually not reliable in realistic scenarios, and overconfidence can be seen as a natural defense against these attacks. Our results suggest that there is a trade-off between reducing a model's overconfidence and its susceptibility to MIAs. Therefore, the informative value of MIAs increases on calibrated models, increasing the privacy risk. As a result, our analysis has shown that MIAs are not as powerful as previously thought and are at odds with the meaning of neural networks' prediction scores.

Acknowledgements

This work was supported by the German Ministry of Education and Research (BMBF) within the framework program "Research for Civil Security" of the German Federal Government, project KISTRA (reference no. 13N15343).

References

[Arora *et al.*, 2018] Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *ICLR*, 2018.

[Choi *et al.*, 2020] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8185–8194, 2020.

[Choquette-Choo *et al.*, 2021] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *ICML*, volume 139, pages 1964–1974, 18–24 Jul 2021.

[Coates *et al.*, 2011] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, volume 15, pages 215–223, 2011.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, volume 70, pages 1321–1330, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hein *et al.*, 2019] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, pages 41–50, 2019.

[Hendrycks and Gimpel, 2017] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[Karras *et al.*, 2020] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020.

[Khosla *et al.*, 2011] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR Workshop*, June 2011.

[Kristiadi *et al.*, 2020] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *ICML*, volume 119, pages 5436–5446, 2020.

[Krizhevsky, 2009] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Master's thesis, Department of Computer Science, University of Toronto, 2009.

[Leibig *et al.*, 2019] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(0), 2019.

[Li and Zhang, 2021] Zheng Li and Yang Zhang. Membership leakage in label-only exposures. In *CCS*, pages 880–895, 2021.

[Müller *et al.*, 2019] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *NeurIPS*, pages 4696–4705, 2019.

[Naeini *et al.*, 2015] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pages 2901–2907, 2015.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*, 2011.

[Nguyen *et al.*, 2015] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015.

[Ramachandran *et al.*, 2018] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *ICLR*, 2018.

[Rezaei and Liu, 2021] Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *CVPR*, pages 7892–7900, 2021.

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

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18, 2017.

[Song and Mittal, 2021] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, 2021.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[Tan and Le, 2019] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, volume 97, pages 6105–6114, 2019.

[Thulasidasan *et al.*, 2019] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(86):2579–2605, 2008.

[Yeom *et al.*, 2018] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE CSF Symposium*, pages 268–282, 2018.