Bachelorarbeit am Institut für Informatik der Freien Universität Berlin,
Arbeitsgruppe ID Management

Group-based Membership Inference Attack against Machine Learning Models

Ina Fendel

ina.fendel@fu-berlin.de

Matrikelnummer:	5205957
Betreuerin:	Dr. Franziska Boenisch
1. Gutachter:	Prof. Dr. Marian Margraf
2. Gutachter:	Prof. Dr. Gerhard Wunder

Berlin, den 18. November 2022

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Berlin, den 18. November 2022

J. Fendel

Ina Fendel

Abstract

Over the last few years, Machine Learning (ML) usage has spread to a wide range of applications, including areas that deal with highly sensitive data. Privacy preservation of data has thus become increasingly important. As ML models have been shown to leak their private training data, an essential part of protecting private data is to avoid privacy leakage of an ML model's private training set.
One important attack risking the privacy of a training set is the membership inference attack (MIA). The MIA identifies an element's *membership*, it analyzes whether a given data point is part of a given model's training set.

This thesis aims at developing new MIAs, especially the in this thesis referred to as *group-based* MIAs. Group-based MIAs determine the membership of an individual data point by exploiting the benefits of grouping elements during the process. For the novel MIAs, methods from the so-called dataset inference attack (DIA) are used. The DIA is a method for ownership resolution which determines whether a model was trained with another model's training set. More importantly for this thesis, the DIA uses a novel method to differentiate between training set elements and other data points, which is applied to this thesis's new MIAs.
This thesis develops four novel MIAs based on the DIA, of which three are group-based, and one does not utilize groups. All approaches were tested on two models trained with CIFAR10 and two further models with CIFAR100 as their training set. The attacks were evaluated with regard to their true positive rate (TPR) at a 0.1 % false positive rate (FPR) and the ROC curve with a log scale, which are metrics found to be suitable for MIAs in previous studies. The results showed that one group-based approach and the not group-based approach work in all settings, while the other two group-based MIAs only work with one of the two tested execution strategies. It was found that the other setup used too ambiguous groups for the attack to work. The experiments further showed that the working group-based MIAs outperform the not group-based approach. The most successful approach overall had a performance of at least 17.9 % TPR at 0.1 % FPR and at best 44.8 % TPR at 0.1 % FPR in the conducted experiments.

Contents

1 Introduction

Machine Learning (ML) has greatly impacted a wide and diverse range of areas, spanning from recommendation systems [30] over e-government [1] to health care [3]. While the usage of ML brings many advantages, it also comes with certain privacy risks [12, 29, 43]. Especially for applications that deal with sensitive data, it is essential to prevent privacy leakage. One important factor is thus attacks against the private training set of an ML model. Understanding these attacks is essential to develop defenses.

In recent years, the membership inference attack (MIA) has emerged as an attack against the concealment of a model's training set. The goal of the attack is to determine whether a given data point is a *member*—in other words, part of the target model's training set [38]. An intuitive example to show its danger is the following: An ML model is trained to find the right medication for cancer patients. Hence, the training data only consists of people with cancer. If an MIA was successfully performed on this model and the used element was indicated to be part of the training set, the adversary would know that the person of the data point has cancer. To hopefully prevent this leakage of highly sensitive data in the future, it is necessary to first research the different possibilities to perform the MIA.

This thesis's goal is to analyze the possibility of a novel MIA based on methods of the so-called dataset inference attack (DIA) [28].

The DIA itself is not an MIA, but builds on the learned knowledge from MIAs that members behave differently to elements outside a target model's training set (called *non-members*) and uses this observation for an attack the other way around: The DIA starts with access to the training set of a model and then applies the MIA principles to reveal whether another model was also trained on this training set. If the other model was indeed trained on the original model's training set, the other model is considered to be stolen from the original model. Put differently, the DIA is used to detect model stealing. While this attack goal itself is not relevant to this thesis, the methods used by DIA are interesting for developing new MIAs. The DIA is performed by calculating how far each data point under test of the original model is away from the decision boundaries to the other classes of the presumably stolen model, under the assumption that members will be further away from the boundaries than non-members. For the DIA, the decision boundary is defined as where, when adapting the data point randomly, its class changes. They call this novel method of walking along the decision boundary Blind Walk. If the model was stolen, it should behave to the members similarly as the original model. The DIA then takes all distances of members from the original model and distances of non-members of the original model to perform a hypothesis test on these two groups. If the statistical p-value is below a certain threshold, the model is confirmed to be stolen. It is explained in the DIA paper [28] that this group-based approach can be more successful at determining whether a point is a member, rather than when focusing on individual points. For this thesis, it is therefore interesting to look at whether it becomes easier to predict the membership if we have more than one data point.

This motivates this thesis's mentioned goal to analyze whether the DIA methods can be used for a new MIA, especially for a group-based MIA.

1.1 Definition of research questions

The following research questions thus have to be looked at:

RQ1 How can the findings and methods of the DIA be used for novel MIAs?

RQ2 Is the DIA's group-based approach applicable for a novel group-based MIA?

RQ3 How well do the novel MIA approaches perform?

1.2 Main contributions

Hence, the main contributions of the thesis are:

- Analysis of the applicability of a group-based approach for a novel MIA. Argumentation about the limitations of the statistical p-value, used in the DIA, for a group-based MIA, and introduction of the effect size as an alternative approach.

- Optimization of the regressor from the DIA to make it applicable for the real-life requirements of an MIA.

- Application of DIA methods for the development of four new MIAs. One novel, not group-based MIA called `threshold-dependent MIA` and three group-based

MIAs called `Removal`, `Growing` and `Replacement` are presented.

- Analysis of the novel MIAs with the result that `Growing`, `threshold-dependent MIA` work without any restraints and `Removal`, `Replacement` in specific settings.

The thesis starts with Chapter 2 about related work, where background information about the MIA and DIA are described. The following methodology-focused Chapter 3 about novel MIA approaches based on the DIA methods is divided into two subsections. The first subsection analyses the applicability of a group-based MIA which is based on the DIA's methods and shows its limitations. In the second subsection, a novel not group-based MIA and novel group-based MIAs are defined, which use the DIA methods.

The implementation details of the new attacks are described in Chapter 4. Chapter 5 presents the results of the MIAs, which are discussed in Chapter 6. The thesis finishes with a conclusion and an outlook in Chapter 7.

2 Preliminaries

The following chapter defines an MIA and names the reasons for its success. This is followed by the related work to this thesis's novel approaches to MIAs by first explaining the already existing group-based MIA called BlindMIA, its success reasons and limitations. The chapter closes with the description of the DIA, which methods will be used in the thesis's attacks, and its success reasons.

2.1 Membership Inference Attack

An MIA is defined as an attack where, given a data point and a target model, the adversary tries to determine the membership of the given data point [38].

In the context of MIAs, a *target model* is an ML model which the adversary attacks. Depending on the attack scenario, the adversary has either white-box or black-box access to the target model. A *member* is a data point that is part of the target model's training set. A *non-member* is a data point that is NOT part of the target model's train set, e.g. a data point from the test set. The *membership* of an element describes whether the element is a member of the given target model.

More formally, a perfectly performed MIA can be defined as (adapted from [31]):

Definition 2.1.1 (Perfect Membership Inference Attack)**.** Let $A : F \times X \to \{1, 0\}$ be a membership inference attack, where F is the set of all ML models and X is the set of all possible input data points for elements in F.
Then:

$$A(f, x) = \begin{cases} 1 \text{ if } x \in D_f \\ 0 \text{ otherwise} \end{cases}$$

where $f \in F$, $x \in X$ and D_f is the set of all data points used for the training of f.

It has to be noted that real-life MIAs do not tend to be perfect MIAs but rather try to come close to being perfect. Accordingly, a real-life MIA might wrongly classify members as non-members and non-members as members. The perfect MIA is solely defined in this thesis to give a better understanding of MIAs and their goal.

MIAs can be performed on a wide range of target models ([7], [35], [36]). This thesis focuses on image classifiers. If not explicitly stated otherwise, image classification is the standard scenario in this thesis.

Further, *group-based* MIAs are mentioned in this thesis. We define this as that the MIA uses group(s) of data instead of individual examples. To avoid confusion, it has to be emphasized that group-based MIAs still have the same goal as other MIAs to determine the membership of individual data points, they just make use of groups in the process of the attack.

2.1.1 Success factors

The reasons why a target model may leak information about whether a given point is an element of its training set, in other words, MIA's success factors, can be categorized as follows:

Overfitting

A model overfits when it has a low training error but fails to generalize well, and as a result has a high test error [14]. The train-test gap (also called generalization error), which calculates the difference between the training and test accuracy of the model [23], makes the impact of overfitting visible: The more a model overfits, the more the training data is ingrained in its structure. When the model then sees the data again, it correctly classifies training data more likely than test data—it has a high train-test gap. Irolla and Chatel examined this relationship: The higher the train-test gap is, the more train data gets correctly classified and the more test data gets misclassified [22].
The MIA builds on differences the model makes between training and test data, making overfitting a success factor of MIAs. Overfitting is a sufficient success reason for MIAs [32, 43], but it is shown not to be necessary for performing an MIA [43].

Target model type and structure

While an overfitted model is more vulnerable to an MIA, its vulnerability highly depends on the model type [42].
A model type describes here which specific kind of ML model is used, e.g., a neural

network or support vector machine (SVM). An overfitted model of type A can still be less vulnerable than a not-overfitted model of type B [38]. This is because a model's type defines how much the decision boundary is impacted by a single element. The more a training point can impact the model structure, the more it is ingrained in the model and the more the model is vulnerable to an MIA [42]. Truex et al. name as an example of a vulnerable model type a decision tree: A member can create a new branch, thereby deeply evolving the structure of the model [42]. This makes it easier for the adversary to see that the element was part of the training set because it has a great impact on the decision boundary and thus becomes easier to differentiate from non-members.

The example shows that not only the type but also an ML model's structure influences a model's vulnerability to MIAs [38].

Training data

There are several success factors for MIAs caused by training data: Limited amount of training data can increase the risk of overfitting, indirectly increasing the vulnerability to MIAs [38]. Shokri et al. have shown that the more training data is available, the weaker the MIA gets [38]. A dataset with a high amount of different classes can also lead to an increased vulnerability because this often results in less training data per class [42]. The less train data available per class, the less uniform the training data is. The ML model learns many small amounts of specialized data, hindering it to generalize well [42]. Little uniformity of data in each class [42] and in general not very representative training data are further success reasons for MIAs [38]. The less uniformity a training dataset has, the more influential a training point is on the decision boundary and the more successful an MIA can be [42].

Difficulty of classification

Difficult classification tasks with an uncertain output are more vulnerable, because the model has to memorize more data than for an easy task [34]. Again, the risk of overfitting is increased, increasing the vulnerability to an MIA [34].
Similarly, tasks with high-dimensional output are more vulnerable to the MIA [34].

2.2 Related Work

This subsection first explains BlindMIA, which is relevant to this thesis's goal because it is an MIA that also works on groups of data points. Secondly, the DIA is described, since this thesis will use DIA's steps and method to develop novel MIA approaches.

2.2.1 BlindMIA

Preliminaries for BlindMIA

The intuition behind BlindMIA is that members behave differently from non-members in a hyper-dimensional space [21].
As a prerequisite to understanding how BlindMIA works, the used hyper-dimensional space (called Reproducing Kernel Hilbert Space (RKHS)) of the attack has to be understood. The following definitions are necessary to comprehend the RKHS:

An inner product space is defined by Szechtman as [40]:

Definition 2.2.1 (Inner Product)**.** An inner product space is a vector space V with an inner product $\langle x\ ,\ y\rangle$ defined on it. An inner product on V is a mapping of $V \times V$ into $\mathbb{R}$ such that for all vectors x, y, z and scalars α, β we have

(i) $\langle \alpha x + \beta y\ ,\ z\rangle = \alpha\langle x\ ,\ z\rangle + \beta\langle y\ ,\ z\rangle$

(ii) $\langle x\ ,\ x\rangle \geq 0$, with equality if and only if $x = 0$.

(iii) $\langle x, y\rangle = \langle y, x\rangle$.

An inner product defines a norm on X given by $||x|| = \sqrt{\langle x\ ,\ x\rangle}$.

The definition of a Cauchy sequence is [8]:

Definition 2.2.2 (Cauchy Sequence)**.** Suppose that $\langle x_n\rangle$ $(n = 1, 2, 3, ...)$ is a sequence of real numbers. $\langle x_n\rangle$ is *Cauchy* if given any $\varepsilon > 0$ there exists an integer $K > 0$ such that $m, n \geq K$ implies $|x_m - x_n| < \varepsilon$.

A Hilbert space is defined in the following way [40]:

Definition 2.2.3 (Hilbert Space)**.** A Hilbert space H is a complete inner product space, complete meaning that every Cauchy sequence in H has a limit in H.

A reproducing kernel is defined as [2]:

Definition 2.2.4 (Reproducing Kernel)**.** A function
$K : E \times E \to \mathbb{C}$
$\quad (s, t) \mapsto K(s, t)$
is a reproducing kernel of the Hilbert space H if and only if:

(i) $\forall t \in E, \;\; K(., t) \in H$

(ii) $\forall t \in E, \;\; \forall \gamma \in H \;\; < \gamma, K(., t) >= \gamma(t)$

(iii) $\forall (s, t) \in E \times E, \;\; K(s, t) =< K(., t), K(., s) >.$

A Hilbert space with a reproducing kernel is called RKHS or proper Hilbert space.

Overview attack

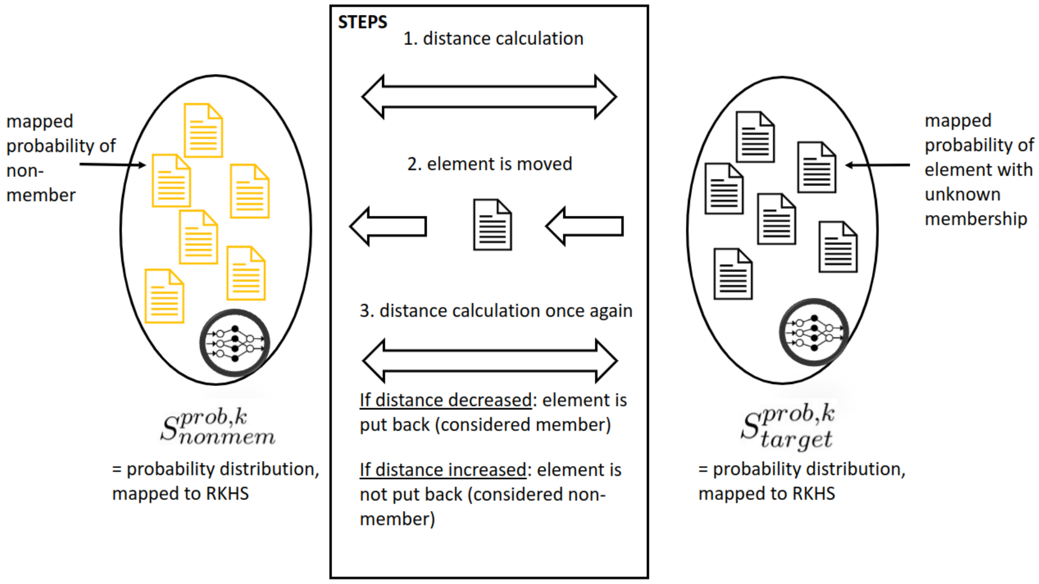

STEPS
1. distance calculation
2. element is moved
3. distance calculation once again
If distance decreased: element is put back (considered member)
If distance increased: element is not put back (considered non-member)
mapped probability of non-member
$S^{prob,k}_{nonmem}$
= probability distribution, mapped to RKHS
mapped probability of element with unknown membership
$S^{prob,k}_{target}$
= probability distribution, mapped to RKHS

Figure 2.1: Visualization of BlindMIA.

A related MIA for image classification tasks with a group-based approach is BlindMIA [21]. BlindMIA is an MIA where the distance between two sets is compared.

The attack is structured in the following way (a visualization of the steps can be found in Figure 2.1):
Two datasets are created, one consisting of non-members and the other one containing elements for which the membership should be determined. The non-members can be created by e.g. the transformation of a given sample or generation of a sample with random features. All elements are run through the target model to get the elements' output probability distributions. The output probability distribution of an element is, for all classes, the target model's predicted probability that the element is from this class. Two new sets are created, one with the non-member output probability distribution, and the other one with the output probabilities of the elements for which the membership has to be determined. The output probability distributions are mapped to the RKHS to make the differentiation of non-members and members easier than in the output probability distribution space.
One set consists now of non-member output probability distributions that are mapped to the RKHS, and the other set of output probability distributions that are mapped to the RKHS of elements for which the membership should be determined. To avoid confusion, the first set is called $S^{prob,k}_{nonmem}$, the second set $S^{prob,k}_{target}$, "prob" standing for output probability distributions, "k" for the k dimensions it got mapped to.
The distance between the two sets is calculated with the following formula [21]:

Definition 2.2.5 ($D(S^{prob,k}_{target}, S^{prob,k}_{nonmem})$)**.** For $y_i \in S^{prob,k}_{nonmem}$, $y_i^{'} \in S^{prob,k}_{target}$, with $S^{prob,k}_{nonmem}$ of size n_n, $S^{prob,k}_{target}$ of size n_t and v is the kernel space dimension, ϕ a feature space map $k \mapsto v$:
$$D(S^{prob,k}_{target}, S^{prob,k}_{nonmem}) = \| \frac{1}{n_t} \sum_{i=1}^{n_t} \phi(y_i) - \frac{1}{n_n} \sum_{j=1}^{n_n} \phi(y_j^{'}) \|_v$$

After the distance between the sets is calculated, one sample gets moved from $S^{prob,k}_{target}$ to $S^{prob,k}_{nonmem}$. The distance between the two sets is calculated again after the move. If the distance between the two sets decreases after the move, the removed example is considered a non-member and stays in the $S^{prob,k}_{nonmem}$. Otherwise, the element is considered a member and moved back in $S^{prob,k}_{target}$.
The comparison of the difference between the distances is called differential comparison. The process is repeated until the distance between the two sets converges. $S^{prob,k}_{nonmem}$ is now considered to only consist of non-members and $S^{prob,k}_{target}$ to mostly, ideally only, consist of members. $S^{prob,k}_{target}$ might be consisting only mostly of members after the distance between the two sets converges, because $S^{prob,k}_{target}$ might still have some non-members whose moving didn't increase the two set's distance. Potential non-members who cause this might be elements that are very similar to members.

Because BlindMIA builds on the comparison of elements with the easily created non-member set, it only needs additionally black-box access to the target model (to get the elements' output probability distributions). The small amount of requirements to perform BlindMIA makes it a widely-applicable attack.

Success factors

BlindMIA's success builds on the observation that the removal or addition of an element influences the position of a whole set in the hyper-dimensional space [21]. Additionally, moving an element from one set to another rather than solely removing the element from one set changes the position of both sets in the hyper-dimensional space, improving the algorithm's sensitivity (sensitivity = true positives (TP)/(TP + false positives (FP)) [11]) [21].

Limitations

The attack is limited by its need for the elements' output probability distributions from the target model. If the adversary has only access to the target model's predicted labels but not the probability distributions, Blind MIA will not work.

Another small limitation is the need to map the output probabilities to RKHS instead of being able to work with the output probabilities directly. This complexity may lead to it being harder to optimize in future works.
A limitation not of the attack itself but its success evaluation is that its evaluation metrics are outdated. Carlini et al. show that the true positive rate (TPR) (TPR = TP/(TP + false negatives (FN))) at a low false positive rate (FPR) (FPR = FP/(FP + true negatives (TN))) is a suitable metric to evaluate MIA, while others can be misleading [6]. Since the BlindMIA paper does not apply TPR at low FPR, it remains unclear how effective the BlindMIA really is.

2.2.2 Dataset Inference Attack

This thesis's MIAs are based on the methods from the DIA [28]. The DIA is not an MIA, it "flips" the MIA to examine whether a complete model was stolen. Model stealing means that without consent of the model's owners, a model's functionality is used for another model [37, 44], a theft of intellectual property [28]. The DIA detects stolen models by exploiting that the training data leaves a signature in the model [28]. Given a model and a training set of another model, it detects whether the model was trained on the given training set. For simplicity, we call the model for which it should be found out whether it was stolen "original model" ($M_{original}$) and the potentially stolen model "suspect model" ($M_{suspect}$).

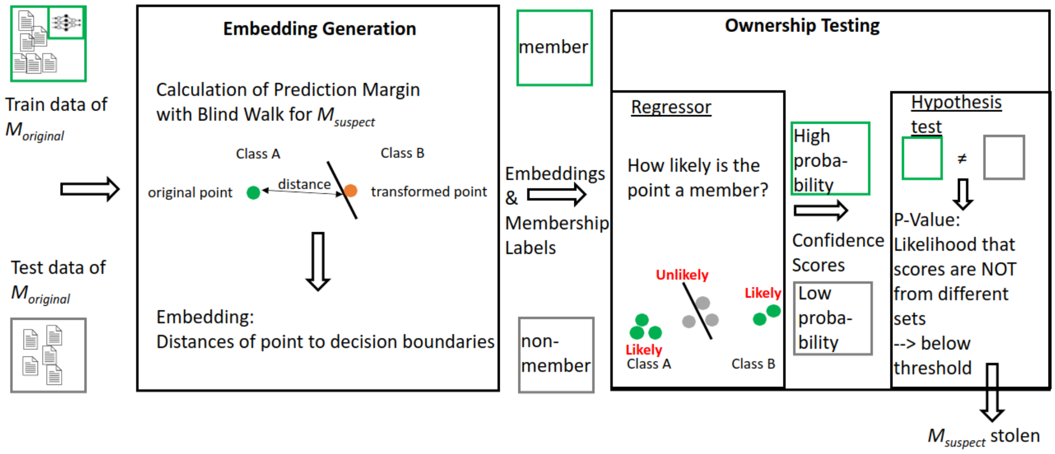

Train data of $M_{original}$
Test data of $M_{original}$
Embedding Generation
Calculation of Prediction Margin with Blind Walk for $M_{suspect}$
Class A
Class B
original point
distance
transformed point
Embedding: Distances of point to decision boundaries
member
Embeddings & Membership Labels
non-member
Ownership Testing
Regressor
How likely is the point a member?
Unlikely
Likely
Likely
Class A
Class B
High proba-bility
Confidence Scores
Low proba-bility
Hypothesis test
≠
P-Value: Likelihood that scores are NOT from different sets --> below threshold
$M_{suspect}$ stolen

Figure 2.2: Visualization of DIA.

Access to the training and test set of $M_{original}$, and black-box access to $M_{suspect}$ are needed to perform the DIA. The attack is divided into two steps, which are visualized in Figure 2.2.

Step 1: Embedding generation

The training and test set of $M_{original}$ are run through $M_{suspect}$ and the points' distances to the decision boundaries to the other classes are calculated. The decision boundary can be explained with an example: If you have an element predicted as one class, and you have changed it enough that the model predicts it as another class, the element has crossed the decision boundary of the model between the two classes. The most successful method presented in the DIA paper to calculate the distances is the novel method Blind Walk, which will be used for this thesis.

Blind Walk

For every data point of the train and test data of $M_{original}$, the following steps are performed:
Random noise is added to the data point. The transformed point runs through $M_{suspect}$. The process is repeated until the predicted class differs from the original point's class. Then, the distance between the original point and the point with a different class is calculated. The distance is considered the prediction margin, which describes the margin of a data point from the decision boundary. To perform Blind Walk, only black-box access to $M_{suspect}$ is necessary.

The vector of distances to the decision boundary for each class created with Blind Walk is called *embedding*. Each embedding gets labeled according to the membership in the original model of the point it was calculated for.

Step 2: Ownership testing

A regressor is trained on the embeddings and their labels. It returns a confidence score for each embedding, indicating how likely the embedding represents a member point. The confidence scores of the train and test set of $M_{original}$ are taken as the input of a hypothesis test. The null hypothesis is that the mean confidence score of the train set is smaller than the test set's mean confidence score. The p-value describes for this hypothesis test how likely the confidence scores are NOT from two different sets. In other words, how likely the suspect model is not stolen. If the p-value is below a set threshold, the null hypothesis will get rejected and $M_{suspect}$ will be considered stolen.

Reason for success

As mentioned as a success reason for MIAs, the DIA also uses the different behavior of a model on train and test data. The DIA assumes that train data has usually a maximized distance to the decision boundary, while test data is generally closer to the decision boundary. Consequently, all stolen models have the original model's train set ingrained in them.

2.2.3 Difference between the DIA and the MIA

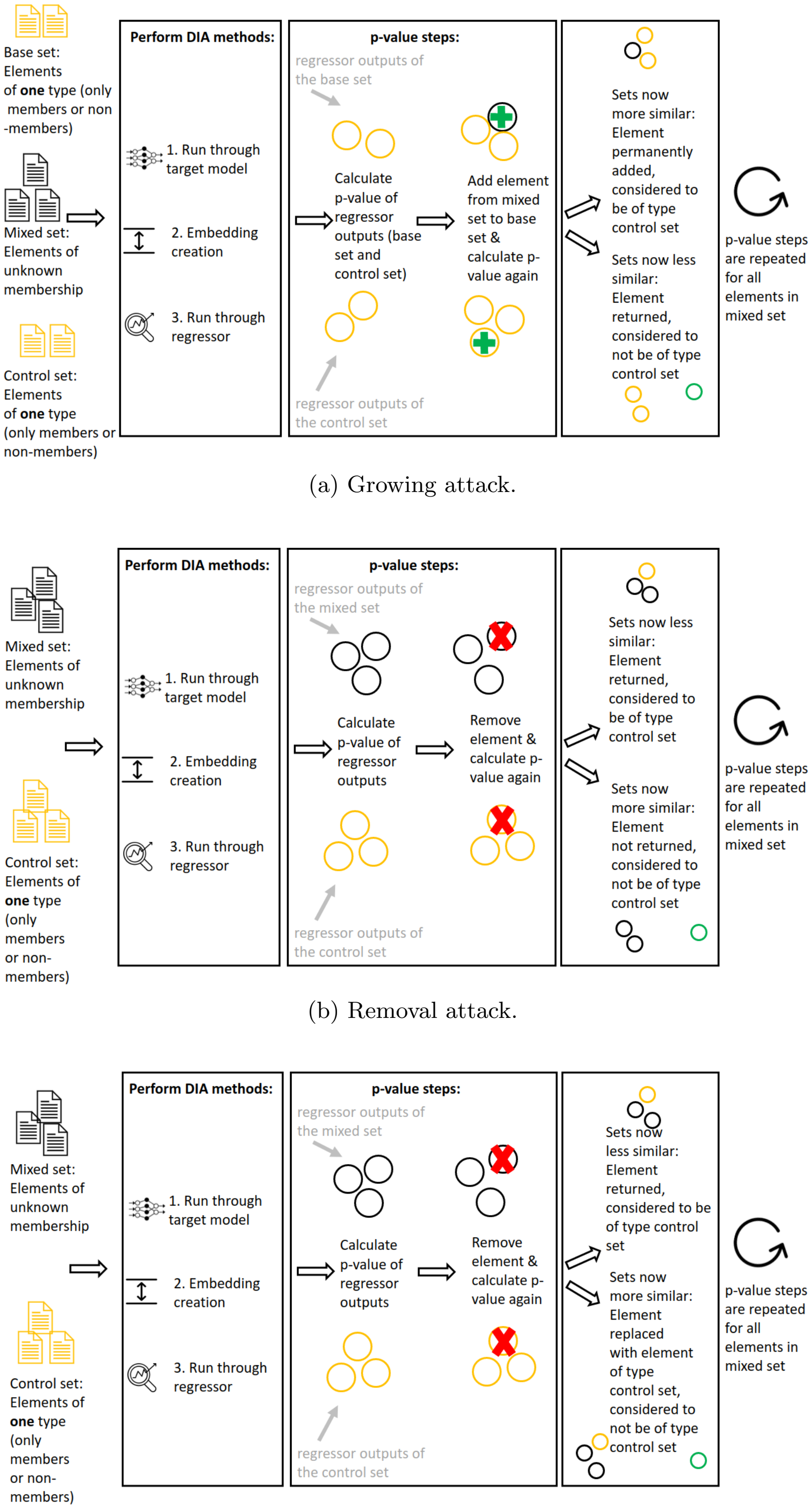
It is important to emphasize that the DIA is not an MIA. Like MIAs, it uses the finding that a model behaves differently on members and non-members, but in contrast to MIAs, where it should be determined whether a given data set has members, the DIA has access to the member set (of $M_{original}$). Additionally, while the purpose of MIAs is to determine the membership of a given dataset for a given target model, the DIA has the different goal of detecting whether a model was stolen.
The motivation behind mentioning DIA in this thesis is that, while it is not an MIA, its methods might be adapted to use for an MIA, especially for a group-based MIA. If a group-based MIA based on the DIA methods succeeded, it would only need access to the predicted labels of the target model instead of the output probability distributions needed for BlindMIA, making it a potentially more versatile attack.

3 Attacks

This chapter focuses on the novel MIAs developed in this thesis. First, the motivation behind MIAs based on the DIA methods is explained. The naive group- and p-value-based attack approaches are defined. p-value limitations for the MIA approaches are then named. Based on these findings, a new not group-based approach without a hypothesis test is proposed, the `threshold-dependent MIA`. This is followed by the explanation of the effect size as an alternative to the p-value for novel group-based MIAs. Finally, three new group-based MIAs, which use the effect size instead of the p-value, are presented (`Removal`, `Replacement`, `Growing`).

3.1 Motivation

This thesis's general focus is to develop a new MIA based on knowledge of the DIA and supported by the approach of the BlindMIA.
The motivation behind this is the following:
The Blind Walk method from the DIA [28] requires very little information to succeed: no additional models have to be trained, only black-box access to the target model is needed and the necessary non-members can be created easily [28]. Blind Walk is therefore a very promising method for a novel MIA.
The DIA combines then the Blind Walk with a regressor and a hypothesis test to statically work on groups. BlindMIA showed a way to perform a group-based MIA through dynamically working with groups. When combining the DIA methods with the dynamic group-based approach of the BlindMIA, it can be used for a new group-based MIA. The novel group-based attack would have the advantage over BlindMIA that it does not need the target model's output probability distribution on the data but only the predicted labels, since only they are necessary for the DIA methods. In contrast to the BlindMIA, this would cause the novel group-based attack to succeed in cases where the adversary would only have access to the target model's predicted labels.


Base set: Elements of **one** type (only members or non-members)
Mixed set: Elements of unknown membership
Control set: Elements of **one** type (only members or non-members)
Perform DIA methods:
1. Run through target model
2. Embedding creation
3. Run through regressor
p-value steps:
regressor outputs of the base set
Calculate p-value of regressor outputs (base set and control set)
Add element from mixed set to base set & calculate p-value again
regressor outputs of the control set
Sets now more similar: Element permanently added, considered to be of type control set
Sets now less similar: Element returned, considered to not be of type control set
p-value steps are repeated for all elements in mixed set
Mixed set: Elements of unknown membership
Control set: Elements of **one** type (only members or non-members)
Perform DIA methods:
1. Run through target model
2. Embedding creation
3. Run through regressor
p-value steps:
regressor outputs of the mixed set
Calculate p-value of regressor outputs
Remove element & calculate p-value again
regressor outputs of the control set
Sets now less similar: Element returned, considered to be of type control set
Sets now more similar: Element not returned, considered to not be of type control set
p-value steps are repeated for all elements in mixed set
Mixed set: Elements of unknown membership
Control set: Elements of **one** type (only members or non-members)
Perform DIA methods:
1. Run through target model
2. Embedding creation
3. Run through regressor
p-value steps:
regressor outputs of the mixed set
Calculate p-value of regressor outputs
Remove element & calculate p-value again
regressor outputs of the control set
Sets now less similar: Element returned, considered to be of type control set
Sets now more similar: Element replaced with element of type control set, considered to not be of type control set
p-value steps are repeated for all elements in mixed set

(a) Growing attack.

(b) Removal attack.

(c) Replacement attack.

Figure 3.1: Naive attack approaches with p-value.

3.2 Naive approaches with p-value

Intuitively, the idea would be to approach the MIAs by applying the DIA methods (including the p-value) straight-forward to the attacks. As explained in the next subchapter, these approaches with p-value have to be discarded as possible attacks because of the p-value limitations. However, to understand why and how, the naive attack approaches first have to be explained in this section. The attacks can be separated into “p-value-based Growing”, “p-value-based Removal” and “p-value-based Replacement” and are visualized in Figure 3.1.

3.2.1 p-value-based Growing

p-value-based Growing starts with a small base set of members. The members are taken in advance, e.g. they can be leaked elements. An element with an unknown membership is added to the set. The DIA methods (embedding generation with Blind Walk, ownership testing with regressor, and hypothesis testing) are applied. The p-values before and after the element is added are compared. Let’s consider the control set to be members, where a member is added to for every element added to the base set.
If the p-value with the added element is higher, the element is considered a member and added to the base set. If the p-value is lower, the element will be considered a non-member and not kept in the set. The process is repeated for all elements with an unknown membership. Through the attack, the base set “grows” in members. In this way, it is possible to fulfill the goal of an MIA to get the membership of the elements with before unknown membership: The members are the added elements in the base set, and the non-members are the removed elements.

3.2.2 p-value-based Removal

p-value-based Removal starts with a mixed set, the membership of the set’s elements are unknown. The same DIA methods as for the p-value-based Growing are applied, and an element is removed. The p-values of before and after the element is removed are compared. Again, the control set is a member set, where members are removed from for every removed element from the mixed set. If the p-value increased, the removed element is considered a non-member, else a member. The process is repeated for the whole mixed set, members are returned to the set, non-members are permanently removed. Step by step, the process “removes” all non-members from the mixed set.

3.2.3 p-value-based Replacement

The third attack is called p-value-based Replacement. Similarly to the p-value-based Removal it starts with a mixed set with elements of unknown memberships, the DIA methods are applied, an element is removed and the p-values before and after the removal are compared. The control set is again a member set. If the p-value increased and therefore the element is considered as a non-member, it is permanently removed from the set, members are put back in. The next step differs from the Removal attack because for the removed non-member a new member is added to the set. The process is repeated for all initial elements in the mixed set. In contrast to p-value-based Growing and p-value-based Removal, the size of the mixed set never changes in p-value-based Replacement. The attack “replaces” all non-members from the mixed set, only keeping the members.

All attacks can similarly be performed with a non-member control set and where non-members are permanently added/ members are permanently removed/replaced. The additional non-members and members for the control sets and replacement have to be created in advance with e.g. a method from the BlindMIA paper [21], other methods to create members [38] or leaked elements are used.

3.2.4 Limitations of p-value-based attacks

A straightforward transformation of the DIA to an MIA includes the hypothesis test and the resulting p-value. An intuitive idea for an MIA based on the DIA would be, as explained above, to compare the p-values.
Normally, hypothesis tests have sets of the same size, as it is the case for the DIA. The p-value is then used to either reject the null hypothesis or to fail to reject the null hypothesis [24]. The described MIA based on the DIA would also perform hypothesis testing on the same-sized sets when calculating the p-value before and after adding/removing/replacing elements. But it would further compare p-values from different hypothesis tests, where the sets of the tests are differently sized: For p-value-based Growing, the hypothesis test before adding elements is performed on smaller sets than after elements are added. For Removal, the sets are bigger for the hypothesis test before elements are removed. In either approach, the calculated p-values that should be compared stem from differently sized samples.

The sample size however impacts the p-value calculation:
Small sample sizes make the calculated p-value less reliable because they less accurately represent their whole set [5]. E.g. a small subset of the target model’s member set represents the member set more poorly than a big subset. The lower reliability

can also lead to irreproducible p-values [19].
A big sample size makes it more likely that the calculated p-value catches a significant difference [4]. But big sets can be unreliable in a different way: Independent of the samples themselves, the calculated p-value will almost always show a significant difference [9, 27, 39].
As a result, the comparison of p-values from differently sized sample sets can be unreliable, since the sample size impacts the p-value calculation.
But even if the sample size is the same for two hypothesis tests (like for p-value-based Replacement), the ability to compare the p-value has its limitations:
Besides the named reasons why a p-value can be unreliable because of the sample size, many more parameters can influence a p-value calculation, e.g. different mean differences and standard errors for two hypothesis tests can result in the same p-value [16]. Further, for the same tested hypothesis, the same p-values from different hypothesis tests do not mean that the results are the same [15, 16]. Gelman and Stern showed that studies compared by their significance level can cause misleading and wrong results [13]. This can be easily explained:
The p-value describes the "probability of the observed result, plus more extreme results, if the null hypothesis were true" [15], so it can only reject or fail to reject the null hypothesis. The p-value itself says nothing explicitly about the alternative hypothesis [15] nor about the magnitude of the difference between the two sets [39]. This is why the comparison of p-values would in this thesis's case compare with the wrong focus, it would be unable to help define whether the added/removed/replaced elements were members.

The theoretical reasoning why the p-value is not applicable for the MIA is backed by the performed experiments. The p-value-based Replacement attack, where the size set does not change and the p-value would therefore be most likely to work, visualizes well its failure. It works as expected up until the last step, the p-value calculation. That the methods before the hypothesis test are working can be seen in Figure 3.2 which displays the sum of regressor outputs of a set before subtracted by after elements were replaced. The expectations for the regressor outputs sums differences are:

- In a member set: Members are replaced by members → they should have similar regressor outputs because they have the same membership → the regressor outputs sums difference should be around zero

- Similar: In a non-member set: Non-members are replaced by non-members → they should have similar regressor outputs because they have the same membership → the regressor outputs sums difference should be around zero

- In a member set: Members are replaced by non-members → they should have dissimilar regressor outputs because they have different memberships → the

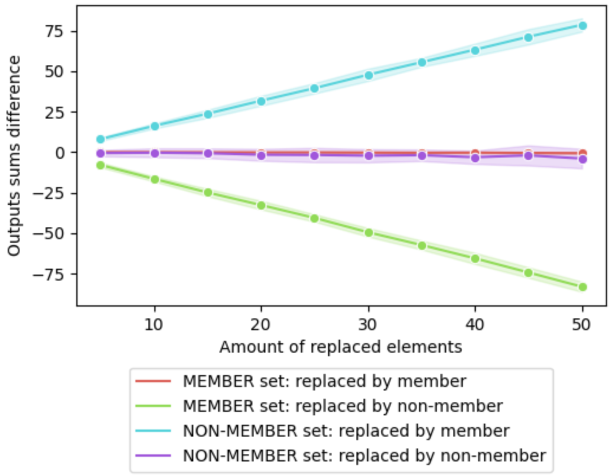


Figure 3.2: Regressor outputs sums differences of sets where specific amounts of elements were replaced. The outputs sums differences on the y-axis describe: The sum of all regressor outputs for the whole set before elements were replaced, subtracted by the sum after elements were replaced.

regressor outputs sums difference should become more and more negative the more members are replaced by non-members (Background information: members have usually a negative regressor output, non-members a positive output)

- Similar, in a non-member set: Non-members are replaced by members → they should have dissimilar regressor outputs because they have different memberships → the regressor outputs sums difference should become more and more positive the more non-members are replaced by members

As seen in Figure 3.2, the experiment matches the expectations, showing that the steps before the hypothesis test worked correctly.
For the p-value calculation, the following is expected:

- In a member set: Members are replaced by members → the set before and after replacement should have similar p-values because the replaced and elements for replacement have the same membership → the p-values difference should be around zero

- Similar: In a non-member set: Non-members are replaced by non-members → the set before and after replacement should have similar p-values because the replaced and elements for replacement have the same membership → the p-values difference should be around zero

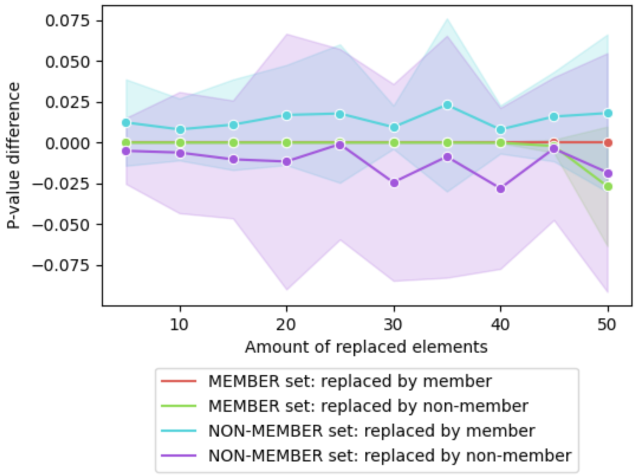


Figure 3.3: p-value differences of sets where specific amounts of elements were replaced (before replacement subtracted by after replacement).

- In a member set: Members are replaced by non-members → the set before and after replacement should have dissimilar p-values because the replaced and elements for replacement have different memberships → the p-values difference should be more and more positive the more members are replaced

- Similar: In a non-member set: Non-members are replaced by members → the set before and after replacement should have dissimilar p-values because the replaced and elements for replacement have different memberships → the p-values difference should be more and more positive the more non-members are replaced

The resulting p-values in Figure 3.3 however do not follow the expectations and seem to be quite random.
To conclude, p-values seem not applicable nor comparable in the context of this thesis. They can be unreliable and misleading in this setting, making the comparison unreliable and misleading as well.

3.3 Alternative to p-value: No hypothesis test

An alternative, not group-based approach to avoid the p-value limitations is to not perform a hypothesis test but to focus on the plain regressor outputs. The group-

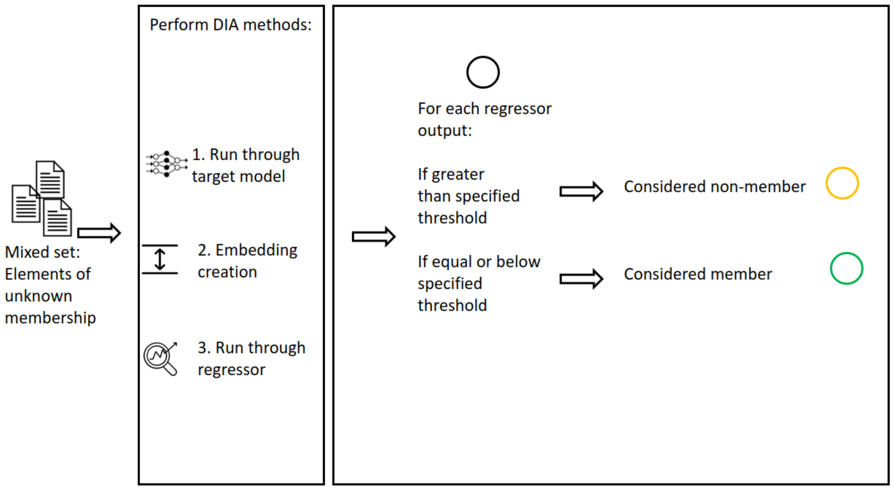

Mixed set: Elements of unknown membership
Perform DIA methods:
1. Run through target model
2. Embedding creation
3. Run through regressor
For each regressor output:
If greater than specified threshold
Considered non-member
If equal or below specified threshold
Considered member

Figure 3.4: `Threshold-dependent MIA`. The specialized threshold has to be determined in advance.

dependent hypothesis test is dropped, and the regressor outputs are determined individually for each element, independently of each other.

3.3.1 Threshold-dependent MIA

The thesis's approach for an MIA based on the regressor outputs is called `threshold-dependent MIA`. It is visualized in Figure 3.4
For the `threshold-dependent MIA`, the elements for which the membership should be determined are the input. Their embeddings are generated with DIA's Blind Walk and are then run through the regressor. If the regressor output is below a set threshold, the element is labeled as a member. Otherwise, the element is considered a non-member. The MIA's output is "dependent" on the threshold.
The regressor has to be trained in advance on members and non-members, and the threshold value has to be determined in advance. Section 4.3 in the next chapter, Chapter 4, will focus on how the threshold can be found.

3.4 Alternative to p-value: Effect size

While a not group-based approach should be sufficient for an MIA to succeed, Blind-MIA [21] showed the promising value of a group-based approach, and the possibility of it being more successful than a non-group-based MIA. This is why in this thesis group-based approaches are further examined besides the presented non-group-based `threshold-dependent MIA`.

Another way than the p-value to compare the difference between two groups is the *effect size* [18, 25, 39].
Effect size is defined as "the magnitude of the difference between two groups" [39]. To clarify the difference between effect size and p-value: p-value can show whether there is a difference between two groups but not the size of the difference [18, 39], showing the size of a difference is what the effect size does.
This makes the effect size suitable to look at in this thesis's context and even superior to the p-value because it focuses on what effect the added/removed/replaced elements have on the set—rather than whether it has an effect as the p-value does. In other words, it answers exactly the question of this thesis's group-based MIA.
Most importantly, it is applicable for the attack in contrast to the p-value because effect sizes are comparable with each other in this thesis's context:
Effect size is independent of the set size [39], which means that when elements are removed or added to the set (as done in the `Removing`, `Growing` attack), this alone does not influence the effect size calculation. A standardized effect size is used in this thesis, which makes it scale-free and the effect sizes comparable across different studies [10]. To sum up, the effect sizes can be compared reliably in contrast to the p-value, making it a suitable alternative for this thesis's group-based MIA.

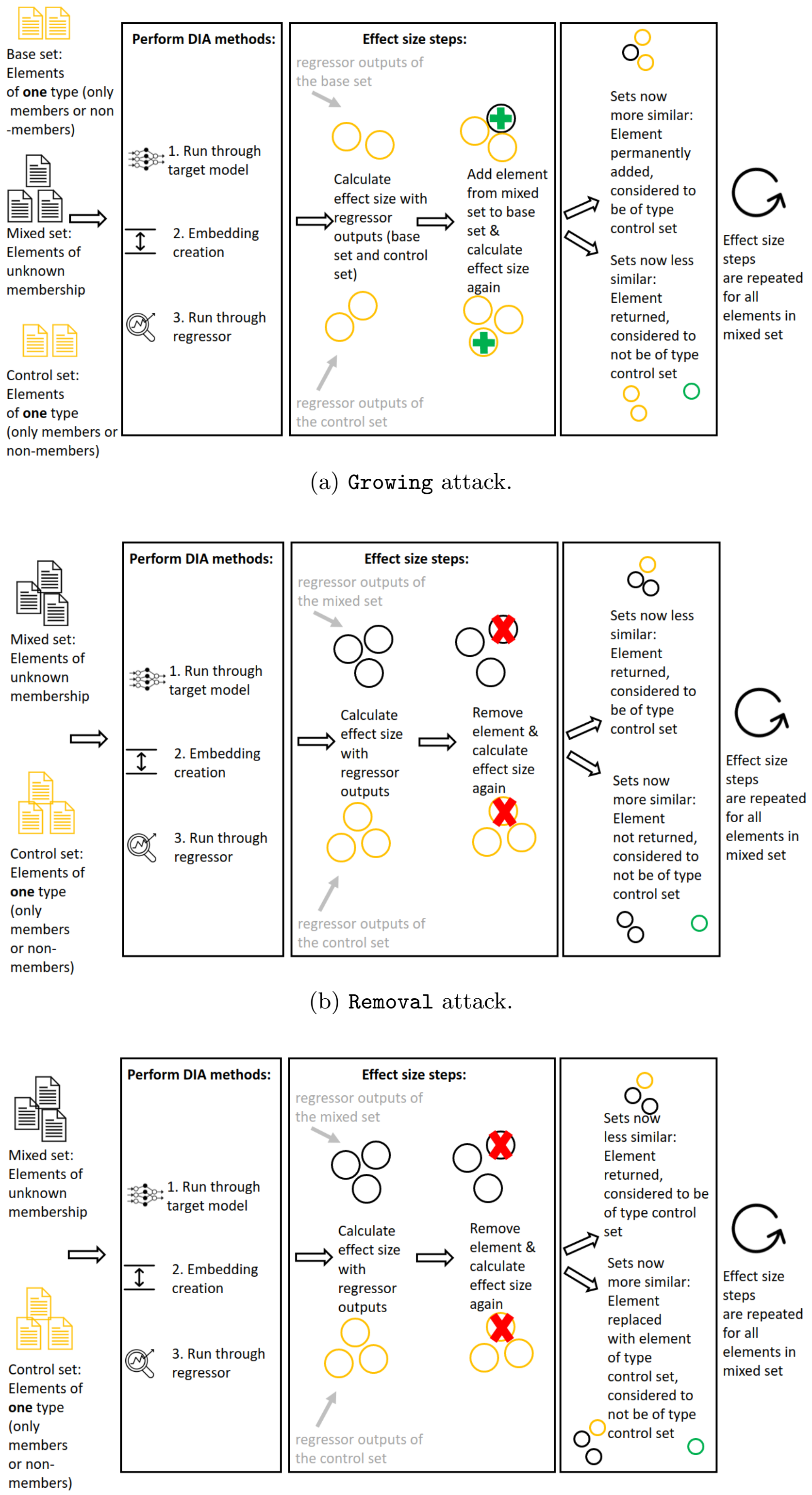
The used effect size is Cohen's d. It is part of the "d-family" of effect sizes, which measure the difference between groups [10], as it is the case in this thesis. Cohen's d assumes that the standard deviation of both groups is roughly the same [10, 25]. Looking at this thesis's experiments, the control group and the set, where elements are removed/added/replaced, have mostly a similar standard deviation because most of the elements are of the same membership. It is important to note that for `Removal` and `Replacement` the standard deviation of the mixed set initially differs from that of its homogenous control set, until the mixed set mainly consists of one type. Nevertheless, experiments for this thesis showed that Cohen's d still works well in these cases, leading to the assumption that the standard deviations are still close enough for Cohen's d to work.
Additionally, normality is mentioned in [26] as a condition for Cohen's d, while not mentioned in other sources [10]. Again, normality is only given for sets where the majority of elements are of one membership, not for an initially mixed set. Still, Cohen's d seems to also work sufficiently well enough in these cases. To conclude, the assumptions for Cohen's d are mostly fulfilled in the experiments, and in the cases where they initially might not, the discrepancy appears to be small enough for Cohen's d to nevertheless work correctly, making it a suitable effect size for this thesis. In preparation for the thesis, it was also found empirically that Cohen's d works the best compared to other effect sizes in the given context.

Cohen's d is defined in the following way [10]:

Definition 3.4.1 (Cohen's d)**.** For M_1 = mean of group one, M_2 = mean of group two,
Cohen's $d = \frac{M_1 - M_2}{SD_{pooled}}$
where SD_{pooled}, the pooled standard deviation, is for two groups A and B of size n_A, n_B and with means $\overline{X_A}$, $\overline{X_B}$
$SD_{pooled} = \sqrt{\frac{\sum(X_A - \overline{X}_A)^2 + \sum(X_B - \overline{X}_B)^2}{n_A + n_B - 2}}$

3.4.1 Group-based MIA with effect size

The group-based attack is performed quite similarly to the discarded p-value-based attack approaches (p-value-based Growing, p-value-based Removal, p-value-based Replacement). The big difference to the discarded p-value-based approaches is that as the last step not the p-value but the effect size gets calculated and compared. The visualization of the attacks can be found in Figure 3.5. The attacks are called `Growing`, `Removal`, and `Replacement`.
For `Growing`, the effect size gets calculated before and after the element with the unknown membership is added to the base set and for `Replacement` before and after an element is replaced.
Similarly, for `Removal` the effect sizes before and after an element is removed from the mixed set are compared.
The other difference to the discarded p-value-based attack approaches is that no elements are removed/replaced/added from/to the control set. It does not change during the attacks, always consisting of as many elements as possible. Effect size does not need the sets to be of the same size, and keeping the biggest possible size for the control set can avoid certain problems that might arise otherwise. For example, if for `Removal` elements were removed from the control set, it would be possible that the control set would become very small, and the kept elements would by chance be not good representations of their membership type, leading to misleading effect sizes.


Base set: Elements of **one** type (only members or non -members)
Mixed set: Elements of unknown membership
Control set: Elements of **one** type (only members or non-members)
Perform DIA methods:
1. Run through target model
2. Embedding creation
3. Run through regressor
Effect size steps:
regressor outputs of the base set
Calculate effect size with regressor outputs (base set and control set)
Add element from mixed set to base set & calculate effect size again
regressor outputs of the control set
Sets now more similar: Element permanently added, considered to be of type control set
Sets now less similar: Element returned, considered to not be of type control set
Effect size steps are repeated for all elements in mixed set

(a) Growing attack.

Mixed set: Elements of unknown membership
Control set: Elements of **one** type (only members or non-members)
Perform DIA methods:
1. Run through target model
2. Embedding creation
3. Run through regressor
Effect size steps:
regressor outputs of the mixed set
Calculate effect size with regressor outputs
Remove element & calculate effect size again
regressor outputs of the control set
Sets now less similar: Element returned, considered to be of type control set
Sets now more similar: Element not returned, considered to not be of type control set
Effect size steps are repeated for all elements in mixed set

(b) Removal attack.

Mixed set: Elements of unknown membership
Control set: Elements of **one** type (only members or non-members)
Perform DIA methods:
1. Run through target model
2. Embedding creation
3. Run through regressor
Effect size steps:
regressor outputs of the mixed set
Calculate effect size with regressor outputs
Remove element & calculate effect size again
regressor outputs of the control set
Sets now less similar: Element returned, considered to be of type control set
Sets now more similar: Element replaced with element of type control set, considered to not be of type control set
Effect size steps are repeated for all elements in mixed set

(c) Replacement attack.

Figure 3.5: Group-based MIAs with effect size named Growing, Removal and Replacement.

4 Implementation

This chapter introduces the implementation details of the attacks presented in the previous Chapter 3. It starts with the description of the parts that are applied for all attacks, e.g. the target models. How the regressor should be optimized for the attacks is explained in the next section. This is followed by implementation details specifically for the `threshold-dependent MIA` and then details specifically for the group-based MIAs. The chapter ends with a description of the used evaluation metrics for the attacks.

4.1 General

The group-based (`Removal`, `Replacement`, `Growing`) MIAs and `threshold-dependent MIA` are all performed in the same setting:
The attacks are executed on the CIFAR10 and CIFAR100 datasets.
The target models for the attacks on CIFAR10 data are a MobileNet [20] and a GoogLeNet model [41]. The MobileNet model has a train accuracy of 94 % and a test accuracy of 89 %. The GoogLeNet model's train accuracy is 98 % and test accuracy is 90 %. For the CIFAR100 attacks, MobileNet [20] and ShuffleNet [45] are used as target models. MobileNet's train accuracy is 90 % and its test accuracy is 67 %. ShuffleNet has a train accuracy of 92 % and a test accuracy of 65 %.
To perform each attack, the adversary needs black-box access to the target model. Black-box access is defined in this thesis similarly to the DIA paper [28], which means that the adversary has only label query access.
Each experiment has a test set of 1000 members and 1000 non-members. To avoid any confusion, the test data is what the novel MIAs (and optimized regressor) are tested on and consists of members (= train data of the MIA's target model) and non-members (= test data of the MIA's target model). When "test set" is mentioned from now on in the thesis, it means the test data of the experiments, not a set of only non-members.

4.2 Membership confidence regressor

The regressor to predict the membership confidence based on the decision boundary distances of an element (= Blind Walk embedding) from the DIA paper [28] is used. Since it has to be trained on members and non-members in advance and especially members as private elements can be hard to retrieve in the context of an MIA, it was optimized to assure the minimal amount of needed members.

4.2.1 Regressor optimization

All training data for the regressor has to be created beforehand. Non-members can be created with the e.g. methods of the BlindMIA [21]. They propose to create new non-members by transforming or adding random noise to an existing non-member, adopting an element from another domain, or creating a sample with random features [21]. Either leaked members can be used for the training set or they have to be created as well, e.g. with one of the methods mentioned by Shokri et al.[38]. They suggest to either add noise to known members, create elements based on statistical knowledge about the population the member set was drawn from, or synthesize members by using elements that the target model has a high confidence on [38]. Please note that since (non-)member creation is not the focus of this thesis, the elements for the attacks will not be manually created but taken from the datasets. Nevertheless, working creation methods have been shown in the above-mentioned papers, which could be applied in a real-life attack.

Because non-members and members have to be created/sourced for the regressor's train set in a real-life attack, it is impractical if unnecessarily many elements are required. An experiment is performed in this thesis to find the minimal regressor's training set size for the regressor to work correctly. The regressor is trained with different amounts of training set sizes and the test accuracies get compared to find the best suitable amount of training data.
The test accuracy is not directly of the regressor itself (since it does not label the data but gives confidence scores), it is rather of a classifier on top of the regressor. This classifier simply divides the elements into members and non-members by separating the regressor confidence scores according to a fixed threshold. While the accuracy is not measured directly on the regressor, it was decided to be the best suitable measurement for the regressor in this thesis's interest: The performance of the regressor is important in the context of how well members and non-members can be divided with the regressor outputs as a base. To simplify the context, it is in this thesis talked about the accuracy of the regressor instead of the accuracy of the classifier on the regressor.

In a real-life attack with a limited amount of data, the test accuracy can be calculated for a growing set of created elements, creating more elements and repeating the experiment as long as the accuracy is insufficient. As soon as it is satisfactory, the attacker can stop creating elements. They should now have only created as few elements as necessary.

Members as private elements are naturally harder to create than non-members, especially when there is very limited knowledge about the target model's training set. The fewer members are necessary for the regressor to work correctly, the easier. Therefore, a second experiment is performed to find the minimal necessary amount of members for the regressor's training set after the experiment to find the necessary amount of training data. The process is similar to the first experiment, but instead of different training set sizes, different member set sizes in the training set are looked at. Additionally, a grid search is performed to find the optimal learning rate and the epochs amount and batch size get optimized.

4.3 Threshold-dependent MIA specific

As in more detail described in Chapter 3, the attack runs the embeddings through the optimized regressor and classifies them dependently on whether they are above a certain threshold. Finding the optimal threshold is thus essential for the success of the `threshold-dependent MIA`.

This thesis applies two ways to find the threshold:
Firstly, because it is known that the test set consists of half members and half non-members, the median of all regressor outputs is taken as the threshold. This approach promises a quite accurate threshold but is only possible to use when the distribution of the set is known. It would thus be applicable for attacks where the amount of members and non-members is roughly known and the attack serves to determine which of the elements exactly are members. The idea could also be applied for attacks where the distribution is not half-half, e.g. if it can be assumed that 3/4 of the set are members, not the median but the regressor output that is bigger than 3/4 of the elements can be taken as threshold (because members usually have smaller regressor outputs than non-members). The distribution could also be manually created when e.g. it is likely that the test elements are all members or at least a vast majority, the same amount of non-members could be added to the set to be able to use the median as the threshold.

The second approach for finding the threshold is for when the adversary does not know the distribution in their test set. Then, the regressor outputs are clustered into two sets and the maximum of the lower cluster is taken as the threshold. Alternatively, the minimum of the higher cluster could be taken. This thesis uses k-means clustering for 1-dimensional data, as presented by Grønlund et al. [17].

4.4 Group-based MIAs specific

The group-based MIAs with effect size, which in detail are explained in Chapter 3, differ from the `threshold-dependent MIA` by the application of effect size on the test set (e.g. mixed set for `Removal`) and the control set before and after an element/elements are replaced/removed/added. Ideally, the effect size varies enough between when only one member is replaced/added/removed compared to a non-member, so the attack can be performed by removing/adding/replacing one element at a time. To find out whether this is the case or whether more elements have to be removed/added/replaced at once, a first experiment is performed for the attack:
For a set where the membership is already known (so not the test set), different amounts of randomly chosen members and non-members are repeatedly removed/added/replaced at once and the mean effect size difference (effect size before - effect size after) is looked at. The minimal amount of elements where there is a clear difference visible in the effect size difference between removed/added/replaced members and non-members is then used for the attack. For example, if the mean effect size difference over the repetitions between one removed member is different from the one of one removed non-member, the attack can be performed by removing only one element at once.
The experiment is not performed on the same elements as the test set, but on a different set to adapt this step to a more real-life attack setting. Here, the adversary would not know the true membership of the test set elements, however, they would know the membership of similar data beforehand for the regressor training. This similar data could then be used for the above-described step. The idea behind the experiment is that since the data behaves similarly to the test set, it should clarify also for the test set how many elements have to be removed/added/replaced at once. Based on the results presented in the next chapter, for the following steps, it is considered that removing/adding/replacing one element at once is enough.

Instead of just differentiating between a positive and negative effect size, a threshold is necessary because the performed experiments have shown that the effect size does not always differentiate between members and non-members of the target model perfectly. For example, when a non-member is removed from the mixed set and the control set is a member set, the effect size might increase in some cases instead of decreasing. In other words, it describes that the mixed set is now less like the control set, while in truth it is more like the control set without the non-member. This could be because the specific non-member is not that easy to differentiate from a member. Nevertheless, in these cases, the effect size only increases a little bit while

when a non-member is removed it increases substantially more than when a member is removed. This makes the removal of a non-member vs. the removal of a member still distinguishable with a reasonable threshold.
For getting the suitable threshold, all elements from the test set are removed/replaced /added from/to the mixed set/base set one at a time and the effect size difference is saved for this element's removal/replacement/addition. Instead of then permanently removing/replacing/adding them according to the threshold as it is done in the attack, they get then added/removed again in this step. As a result, each element gets removed/replaced/added from the mixed set (or for `Growing` base set) that the attacks first starts with.
Then, the effect size differences are used to calculate the threshold. The same methods as for the `threshold-dependent MIA` are used: For test sets like the one used in the thesis where the distribution of half members, half non-members is known, the median of the effect size differences is taken as the threshold. For when the distribution is unknown, k-means clustering for 1-dimensional data [17] with k = 2 is applied to the effect size differences and the maximum of the lower cluster is used as the threshold.
The idea behind this step is that the effect size difference should be the least extreme at the beginning of the group-based MIAs. For example, when you remove an element from a mixed set that is half full of members and half full of non-members and calculate the effect size difference compared to a non-member set, the removed element should not have had a big impact. In contrast, later in the attack, when the mixed set consists mostly of non-members, a removed member should have a bigger impact and the effect size difference should be more extreme. As a result, the calculated effect size differences for members and non-members in this step should be the closest to each other that they will ever be in the attack. Calculating the threshold on them should thus result in a good estimation of the optimal threshold for the whole attack.

The actual attack is performed with either member or non-member as the control set. Accordingly, either members or non-members are removed/added/replaced. The control set consists of 1000 elements.
For `Growing`, the base set consists of 500 elements.

4.5 Metrics

This thesis's evaluation of the novel MIAs is based on the findings of Carlini et al. [6], that the ROC curve with log-scale and the TPR at a low FPR are good metrics to evaluate the success of an MIA. The paper proposes to look at TPR at 0.1 % and 0.001 % FPR. Since an FPR of 0.001 % would need a test set of at least size 100 000 elements (1/100 000 = 0.00001 = 0.001 %), while the CIFAR datasets have only

60 000 elements, 0.001 % FPR is disregarded in this thesis. Additionally, the ROC curve without log scale and AUC are looked at to evaluate whether the attack is generally working or random. The amounts of correctly classified members and non-members from the confusion matrix are reported to evaluate how well the threshold is working. The accuracy and F1-score are also reported for completeness, similar to as done with the accuracy by Carlini et al. [6]. Likewise, they will be disregarded in the evaluation because according to [6] they are not informative about the MIAs success.

5 Results

The chapter analyses first the results of the regressor optimization, followed by the results of the `threshold-dependent MIA` and group-based MIAs (`Removal`, `Growing`, `Replacement` with non-member or member control set). “Results” finishes with a summary of the findings.

5.1 Regressor optimization

	DIA regressor for target models trained on CIFAR10		Optimized regressor for target models trained on CIFAR10	
Target model	MobileNet	GoogLeNet	MobileNet	GoogLeNet
Test accuracy	69.40 %	68.80 %	74.90 %	76.40 %
Amount train data	10000		5000	
Amount members in train data	5000		2000	
Amount non-members in train data	5000		3000	
Loss function	DIA loss		MSE loss	
Learning rate	0.1		0.05	
Epochs	1000		400	
Batch size	All elements at once		16	

Table 5.1: Comparison of regressor from DIA paper [28] and optimized regressor for target models trained on CIFAR10.

As explained in Chapter 4, the regressor that predicts membership confidences based on the decision boundary distances was originally taken from the DIA paper [28], and then optimized to better fit the purpose of MIAs. Table 5.1 shows the hyperparameter and performance of the original and optimized regressor for target models trained with CIFAR10, Table 5.2 for target models trained with CIFAR100. For all models, the amount of training data was able to be reduced from 10000 to 5000 elements: Only 2000 members and 3000 non-members in the training set instead of the original 5000 each were necessary for the optimized regressor. The loss function was changed from their own *DIA loss*, which took the mean of the non-squared values, to Mean Squared Error (MSE) for all regressors. MSE is defined as [33]:

	DIA regressor for target models trained on CIFAR100		Optimized regressor for target models trained on CIFAR100	
Target model	MobileNet	ShuffleNet	MobileNet	ShuffleNet
Test accuracy	86.30 %	90.90 %	90.10 %	92.60 %
Amount train data	10000		5000	
Amount members in train data	5000		2000	
Amount non-members in train data	5000		3000	
Loss function	DIA loss		MSE loss	
Learning rate	0.1		0.1	
Epochs	1000		400	
Batch size	All elements at once		16	

Table 5.2: Comparison of regressor from DIA paper [28] and optimized regressor for target models trained on CIFAR100.

Definition 5.1.1 (Mean Squared Error)**.** MSE $= \frac{\sum_{i=1}^{n}(y_i - \lambda(x_i))^2}{n}$
where y_i is the true target value for test instance x_i, $\lambda(x_i)$ is the predicted target value for test instance x_i, and n is the number of test instances.

In comparison, MSE puts larger weight on big errors and punishes small errors less than the DIA loss. While the original regressors put all elements in a single batch, the optimized regressors have a batch size of 16. The epochs were reduced to 400 from 1000 based on when the test accuracy stopped getting better.
For the regressors based on the CIFAR10 dataset, the learning rate was adapted from 0.1 to 0.05 to make the training more stable.
The performances of all regressors were improved by the optimization: The test accuracy of the regressors for CIFAR100 target models improved by 3.8 % (MobileNet) and 1.7 % (ShuffleNet) and for CIFAR10 target models by 5.5% (MobileNet) and 7.6 % (GoogLeNet).
With test accuracies above 90 % for both CIFAR100 in contrast to test accuracies above 70 % for CIFAR10, the optimized regressors work better for CIFAR100 target models.

5.2 Method to determine thresholds

As mentioned in Chapter 4, the threshold for the regressor outputs (`threshold-dependent MIA`) and for the effect size differences (group-based MIAs) were calculated with the median and through clustering. Since the median should be often applicable (and to keep the result chapter length reasonable) the analysis is focused solely on the results that are using the median-based thresholds. The results for the attacks with a threshold through clustering can be found in the appendix. The attacks with a median-based threshold are usually more successful than those with a cluster-based threshold, for both the `threshold-dependent MIA` and group-based MIAs.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Threshold	**Median as threshold**			
Threshold-dependent MIA	0.83654	0.67631	-0.55246	0.38511

Table 5.3: Optimal, rounded thresholds for `threshold-dependent MIA`.

5.2.1 Threshold-dependent MIA

The optimal threshold for classifying the regressor outputs as members or non-members can be found in Figure 5.3.

The performance details of the `threshold-dependent MIA` on the models are presented in Table 5.4 and the ROC curves in Figure 5.1. As seen in Figure 5.1, the attack proves to be successful against all models as it is better than the random baseline.
All attacks have an AUC higher than 0.8, with the attack on the CIFAR100 ShuffleNet model having the highest AUC (0.86). For the MIA on CIFAR10 MobileNet and GoogLeNet, no FPR at 0.1 % was found (5.4). The reason for this is that many elements had the same value as the highest regressor output so the FPR could never become 0.1 %. The closest higher FPRs were 1.1 % for MobileNet and 3.1 % for GoogLeNet, values considerably higher than 0.1 % and therefore hard to use for comparison. At 0 % FPR, the TPR is 0 % for both models. The attacks on the CIFAR100 models had a TPR of higher than 0 % at 0.1 % FPR. The attack on CIFAR100 ShuffleNet has the highest TPR at 0.1 % FPR (6.9 %).
For all median-based thresholds, between 732 and 806 of the 1000 members/non-members are correctly classified. With a difference of 19 (CIFAR10 MobileNet), 2 (CIFAR10 GoogLeNet), 13 (CIFAR100 MobileNet) and 5 (CIFAR100 ShuffleNet) the attacks on all models classify more non-members correctly than members.

5.3 Group-based MIAs with effect size

For all group-based MIAs with effect size, the first pre-experiment before the attack (finding the minimal amount of elements that have to be added/removed/replaced

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Threshold	**Median as threshold**			
TPR at 0.1 % FPR	Not found, 0.1 % FPR does not exist. TPRs at closest existing FPRs: At 1.1 % FPR: 11.7% At 0 % FPR: 0%	Not found, 0.1 % FPR does not exist. TPRs at closest existing FPRs: At 3.1 % FPR: 14.3% At 0 % FPR: 0%	2.2%	6.9%
AUC	0.81	0.84	0.83	0.86
Accuracy	76.25%	80.5%	73.85%	78.65%
F1-score	76.02%	80.48%	73.67%	78.59%
Amount correctly classified members (out of 1000 members)	753	804	732	784
Amount correctly classified non-members (out of 1000 non-members)	772	806	745	789

Table 5.4: Performance of the `threshold-dependent MIA`. When 0.1 % FPR does not exist, the closest FPRs are mentioned.

at once) showed, that one element is enough for the attacks on the target models to divide between member and non-member. Therefore, in the following experiments, only one element is removed/added/replaced at once.

The optimal thresholds from the second pre-experiment i.e., analyzing which effect size difference threshold is suitable to divide between members and non-members, can be found in Table 5.5.

5.3.1 Removal with control set non-members

Figure 5.2 depicts the ROC curves for the `Removal` attack with a non-member control set. All attacks have an AUC of around 0.8, for GoogLeNet and ShuffleNet, the AUC is the highest at 0.83. As it becomes especially visible in Plot (b) of Figure 5.2 (ROC curve with a log scale), all models perform very poorly at low FPRs (see

Model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Datatset	**CIFAR10**		**CIFAR100**	
Threshold type	**Median as threshold**			
Removal (non-member control set)	-8.34E-05	-7.55E-05	0.000408511749173	0.000127381497711
Removal (member control set)	-0.000224209779989	-0.000219135926477	0.000262594563277	-5.91E-05
Growing (non-member control set)	-1.53E-03	-1.32E-03	-0.001269073438641	-0.002302819265117
Growing (member control set)	0.001288472187745	0.001383560564791	4.47E-06	1.03E-03
Replacement (non-member control set)	-7.97E-05	-7.38E-05	0.000408511749173	0.000153541643544
Replacement (member control set)	-0.000224873464333	-0.000264211535473	0.000256300559011	-7.83E-05

Table 5.5: Optimal thresholds for group-based attacks.

Table 5.6). This is represented again in the TPR at 0.1 % FPR, where the attack has on all models a TPR of 0 %. By this metric, the attack fails to work correctly as an MIA.
Despite the overall poor performance, the MIAs work well on members with the medians as thresholds, with at least 902 and up to 929 correctly classified members. Between 680 and 851 non-members are correctly classified, therefore all attacks classify more members correctly (difference of 95 for CIFAR100 MobileNet, 78 for CIFAR100 ShuffleNet, 222 for CIFAR10 MobileNet, 153 for CIFAR10 GoogLeNet).

5.3.2 Removal with control set members

All `Removal` attacks with a control set of members have an AUC of 0.97 (CIFAR10 MobileNet) or 0.98 (other models), as seen in their ROC curves 5.3. All attacks are above the random baseline and thus work. The TPR at 0.1 % FPR, see Table 5.7, is for all `Removal` attacks high (lowest: 17.9 % for CIFAR10 MobileNet, highest: 44.8 % for CIFAR100 ShuffleNet). According to this metric, the attack performs very

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Non-member**			
Threshold	**Median as threshold**			
TPR at 0.1 % FPR	0.00%	0.00%	0.00%	0.00%
AUC	0.79	0.83	0.8	0.83
Accuracy	79.10%	84.65%	86.75%	89.00%
F1-score	81.18%	85.74%	87.35%	89.41%
Amount correctly classified members (out of 1000 members)	902	923	915	929
Amount correctly classified non-members (out of 1000 non-members)	680	770	820	851

Table 5.6: Performance of the `Removal` attack with a control set of non-members.

well and the best on the ShuffleNet model.
For the specific medians as thresholds, between 876 and 937 elements of each type are correctly classified. The MIAs on all models classify more non-members correctly (difference: 47 (CIFAR10 MobileNet), 38 (CIFAR10 GoogLeNet), 21 (CIFAR100 MobileNet), 20 (CIFAR100 ShuffleNet)).

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Member**			
Threshold	**Median as threshold**			
TPR at 0.1 % FPR	17.90%	29.50%	28.10%	44.80%
AUC	0.97	0.98	0.98	0.98
Accuracy	89.95%	92.00%	90.95%	92.70%
F1-score	89.71%	91.85%	90.85%	92.62%
Amount correctly classified members (out of 1000 members)	876	901	899	917
Amount correctly classified non-members (out of 1000 non-members)	923	939	920	937

Table 5.7: Performance of the `Removal` attack with a control set of members.

5.3.3 Growing with control set non-members

Table 5.8 and Figure 5.4 show the performance of the `Growing` attack with a non-member control set. The ROC curves (5.4) show that the attacks are working since they are above the random baseline. All `Growing` attacks have an AUC of 0.92 (CI-

FAR100 MobileNet) or higher. The attack on CIFAR100 MobileNet has the lowest TPR at 0.1 % FPR (0.8 %), the other MIA's rates are considerably higher, CIFAR10 GoogLeNet having the highest TPR (11.5 %).
Besides the amount of correctly classified members at the MIA on the CIFAR100 MobileNet (538), more than 830 elements are at least correctly classified of each type with the medians as thresholds. All attacks classify more non-members correctly: CIFAR10 MobileNet difference is 80, CIFAR10 GoogLeNet is 43, CIFAR100 MobileNet is 407 and CIFAR100 ShuffleNet is 32. With only slightly more than half of the members correctly classified for the CIFAR100 MobileNet model, the attack does not work significantly better than random in this case.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Non-member**			
Threshold	**Median as threshold**			
TPR at 0.1 % FPR	5.60%	11.50%	0.80%	3.00%
AUC	0.93	0.95	0.92	0.95
Accuracy	87.30%	90.25%	74.15%	92.00%
F1-score	86.77%	90.04%	67.55%	91.87%
Amount correctly classified members (out of 1000 members)	833	881	538	904
Amount correctly classified non-members (out of 1000 non-members)	913	924	945	936

Table 5.8: Performance of the `Growing` attack with a control set of non-members.

5.3.4 Growing with control set members

All `Growing` attacks with a control set of members work since they are above the random baseline (see ROC curves 5.5). The CIFAR100 ShuffleNet model attack has the biggest AUC (0.95), all other AUCs are only slightly smaller with 0.93 or higher. The MIA on the GoogLeNet model has also the highest TPR at 0.1 % FPR (see performance attack: 5.9, 11.7 %). The attack on the CIFAR100 MobileNet model has with 0.8 % the lowest TPR at 0.1 % FPR.
658 non-members are correctly classified with the median as the threshold for the CIFAR10 MobileNet as the target model. Besides that, all attacks classify at least above 790 elements of one type correctly with the medians as thresholds. Besides the CIFAR100 MobileNet attack (difference 21), all attacks classify more non-members correctly (CIFAR10 MobileNet: difference 274, CIFAR10 GoogLeNet difference 135, CIFAR100 ShuffleNet distance 8).

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Member**			
Threshold	**Median as threshold**			
TPR at 0.1 % FPR	5.40%	11.70%	0.80%	3.10%
AUC	0.93	0.94	0.93	0.95
Accuracy	79.50%	86.15%	89.95%	91.60%
F1-score	81.97%	87.03%	89.84%	91.63%
Amount correctly classified members (out of 1000 members)	932	929	889	920
Amount correctly classified non-members (out of 1000 non-members)	658	794	910	912

Table 5.9: Performance of the `Growing` attack with a control set of members.

5.3.5 Replacement attack with control set non-members

Table 5.10 and Figure 5.6 present the performance of the `Replacement` attack with a control set of non-members. The AUC is the highest for the attack on CIFAR GoogLeNet (0.81) and the lowest for on CIFAR100 MobileNet (0.2). The ROC curves with log scale (plot (b), 5.6) show that the attack does not work since it is not better than the random baseline. Accordingly, the attacks have a low or even 0 % (CIFAR100 MobileNet) TPR at 0.1 % FPR.
While the medians as thresholds perform quite well on attacks on the CIFAR10 models (at least 763 correctly classified elements of each type), they do not work on the CIFAR100 models, with for ShuffleNet even only 1 member correctly classified. Accordingly, the difference between correctly classified members and non-members is high for CIFAR100 models with 887 (CIFAR100 MobileNet) and 999 (CIFAR100 ShuffleNet). The attacks on CIFAR10 models both classify more members correctly (Difference MobileNet: 17, ShuffleNet: 125).

5.3.6 Replacement attack with control set members

The `Replacement` attack with members as a control set works not randomly since it is above the random baseline (ROC curves, Figure 5.7). The CIFAR100 MobileNet attack has the highest AUC (0.95), and the CIFAR10 MobileNet has the lowest (0.87). The attack on CIFAR100 MobileNet has the highest (13.4 %) TPR at 0.1 % FPR, CIFAR10 GoogLeNet the lowest (5.6 %).
Besides the attack on the CIFAR100 MobileNet model (908 correctly classified members, 898 correctly classified non-members), the medians as thresholds do not work for the `Replacement` attack: For the remaining models, zero non-members are cor-

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Non-member**			
Threshold	**Median as threshold**			
TPR at 0.1 % FPR	0.20%	0.20%	0.00%	1.40%
AUC	0.79	0.81	0.2	0.96
Accuracy	80.15%	82.55%	51.55%	50.05%
F1-score	80.32%	83.57%	66.43%	0.19%
Amount correctly classified members (out of 1000 members)	810	888	959	1
Amount correctly classified non-members (out of 1000 non-members)	793	763	72	1000

Table 5.10: Performance of the `Replacement` attack with a control set of non-members.

rectly classified.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Member**			
Threshold	**Median as threshold**			
TPR at 0.1 % FPR	10.90%	5.60%	13.40%	11.70%
AUC	0.87	0.91	0.95	0.86
Accuracy	49.15%	49.30%	90.30%	48.20%
F1-score	65.90%	66.04%	90.35%	65.05%
Amount correctly classified members (out of 1000 members)	983	986	908	964
Amount correctly classified non-members (out of 1000 non-members)	0	0	898	0

Table 5.11: Performance of the `Replacement` attack with a control set of members.

5.4 Summary of results

Summarizing the results from above, the following statements can be made:

Regressor The regressors were successfully improved while achieving the goal of reducing the amount of train data and members in the train data.

Attacks The `threshold-dependent MIA`, `Removal` with a member control set, `Growing` with a non-member and member control set and `Replacement` with a member control set succeeded as attacks and according to the stricter metrics of [6] as MIAs. Concerning the TPR at 0.1 % FPR, the `Removal` with member control set was overall the most successful MIA.

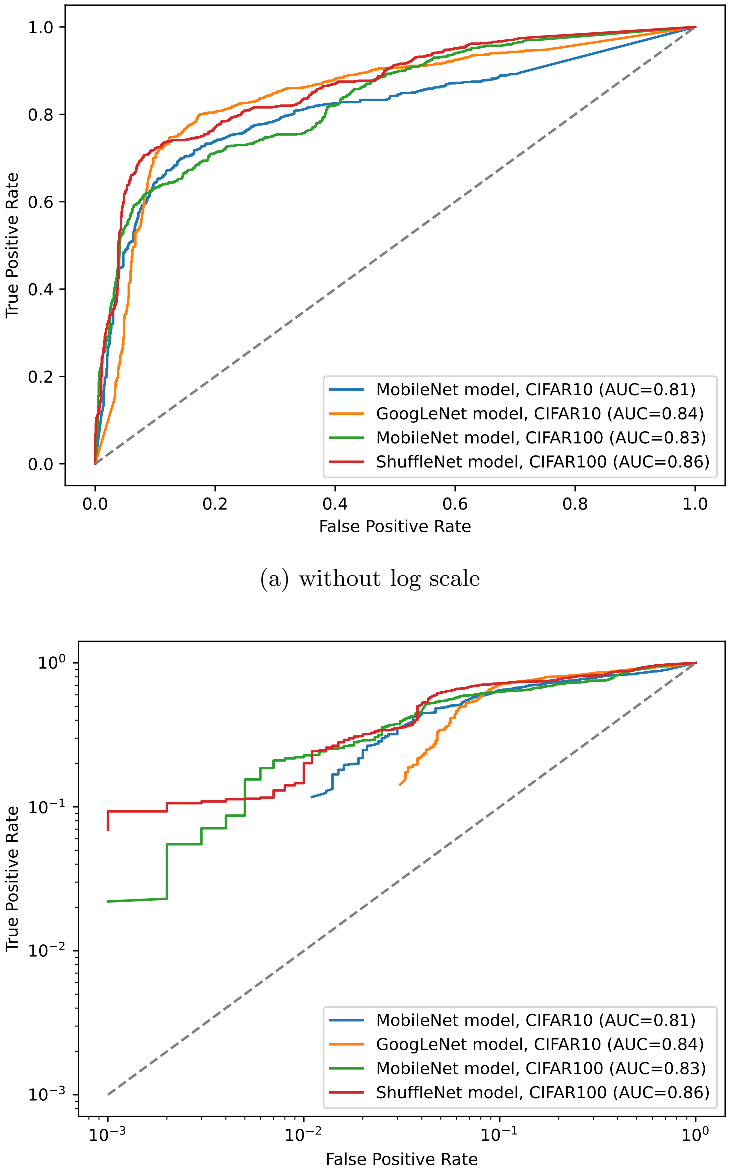
The attack `Removal` with a non-member control set failed to work and `Replacement` with a non-member control set only partially worked as an attack.

The `threshold-dependent MIA` performed better on the CIFAR100 target models than on the CIFAR10 target models.

For the group-based attacks, no clear pattern emerges about the success of the attacks depending on the target models. CIFAR100 ShuffleNet and CIFAR10 GoogLeNet are in two attacks the target models where the attack is most successful on, CIFAR100 MobileNet one time. The MIAs work the worst three times on the CIFAR100 MobileNet model and one time on CIFAR10 MobileNet and CIFAR10 GoogleNet each.

Medians as thresholds for attacks The MIAs with medians as their thresholds for the `threshold-dependent MIA`, `Removal`, and `Growing` attacks worked (besides for MobileNet 100 for `Growing` with non-member control set). The `threshold-dependent MIA`, `Removal` with a member control set, and `Growing` with a non-member control set all classify slightly more non-members than members correctly. For `Removal` with a non-member set and `Growing` with a member set (besides CIFAR100 MobileNet), slightly more members are correctly classified.

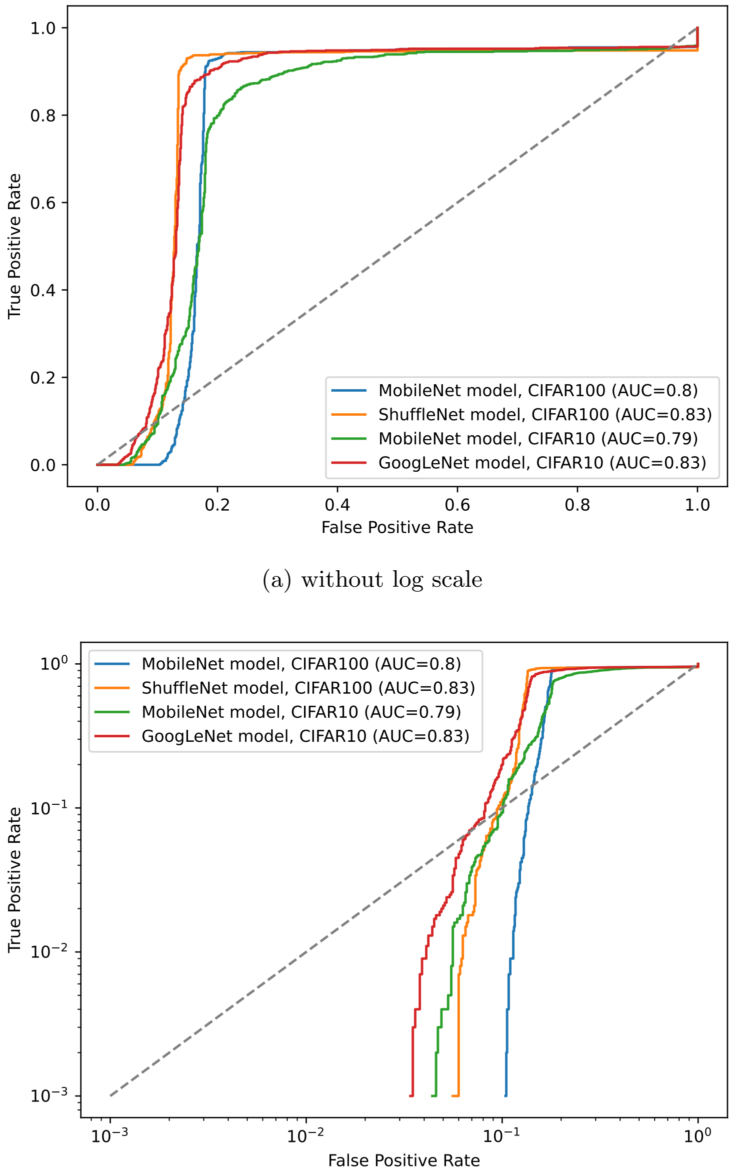
The thresholds failed to separate the elements correctly for the `Replacement` attacks (besides for CIFAR10 models for `Replacement` with non-member control set and CIFAR100 MobileNet for `Replacement` with member control set).



(a) without log scale

(b) with log scale

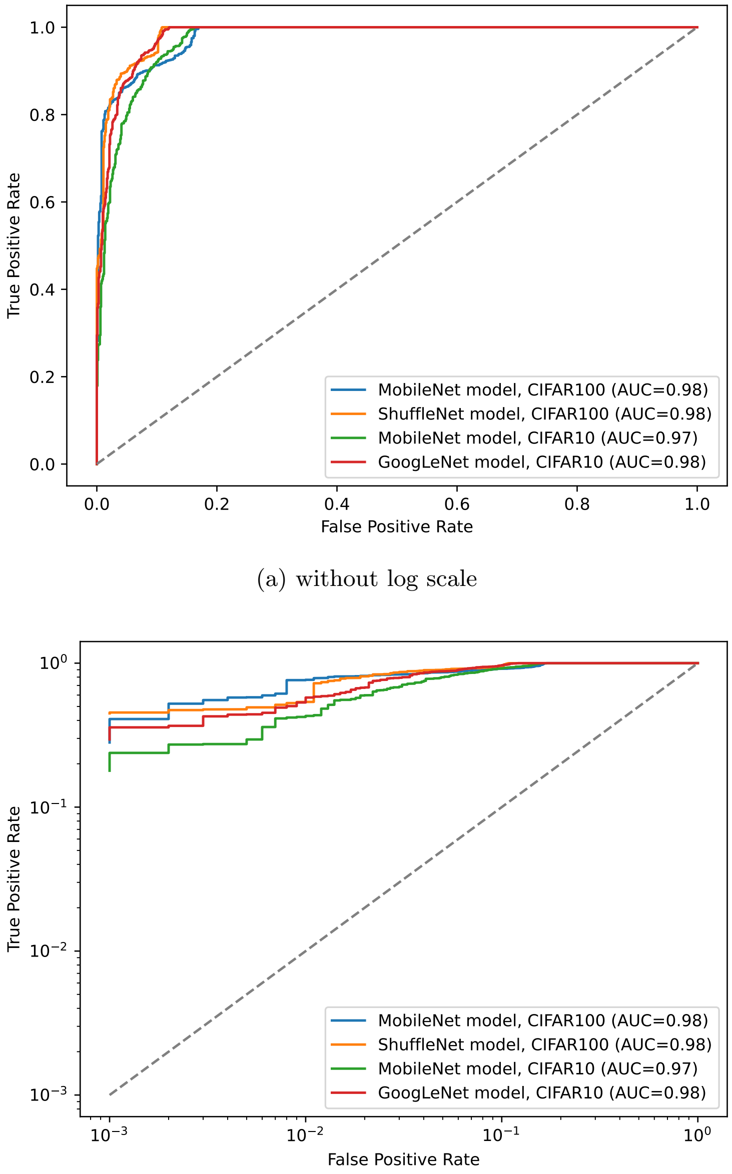
Figure 5.1: ROC curve without log scale and with log scale for the `threshold-dependent MIA`.



(a) without log scale

(b) with log scale

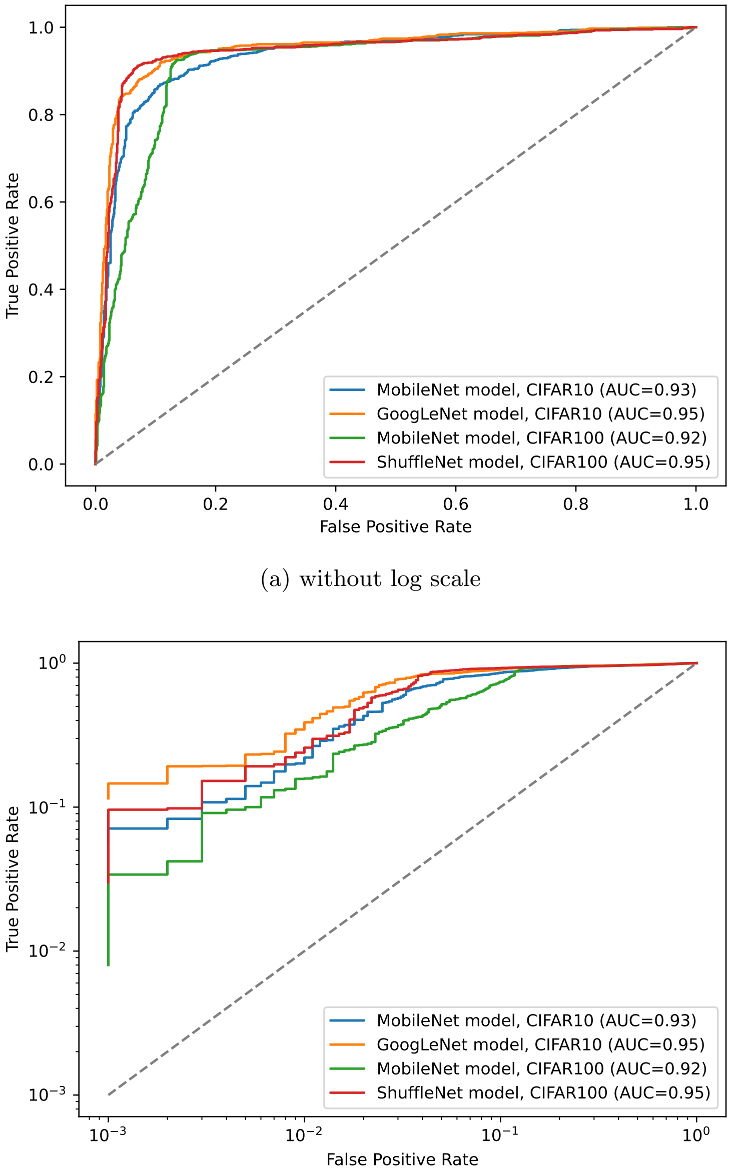
Figure 5.2: ROC curve without log scale and with log scale for the `Removal` attack with a control set of non-members.



(a) without log scale

(b) with log scale

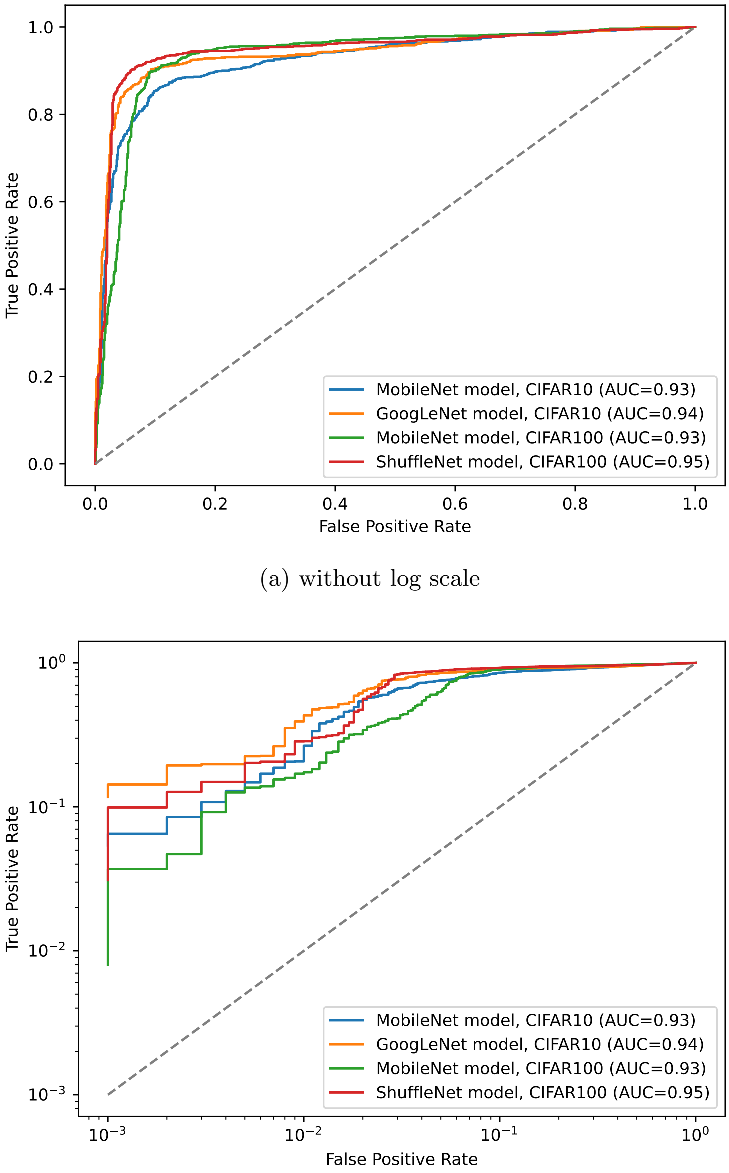
Figure 5.3: ROC curve without log scale and with log scale for the `Removal` attack with a control set of members.



(a) without log scale

(b) with log scale

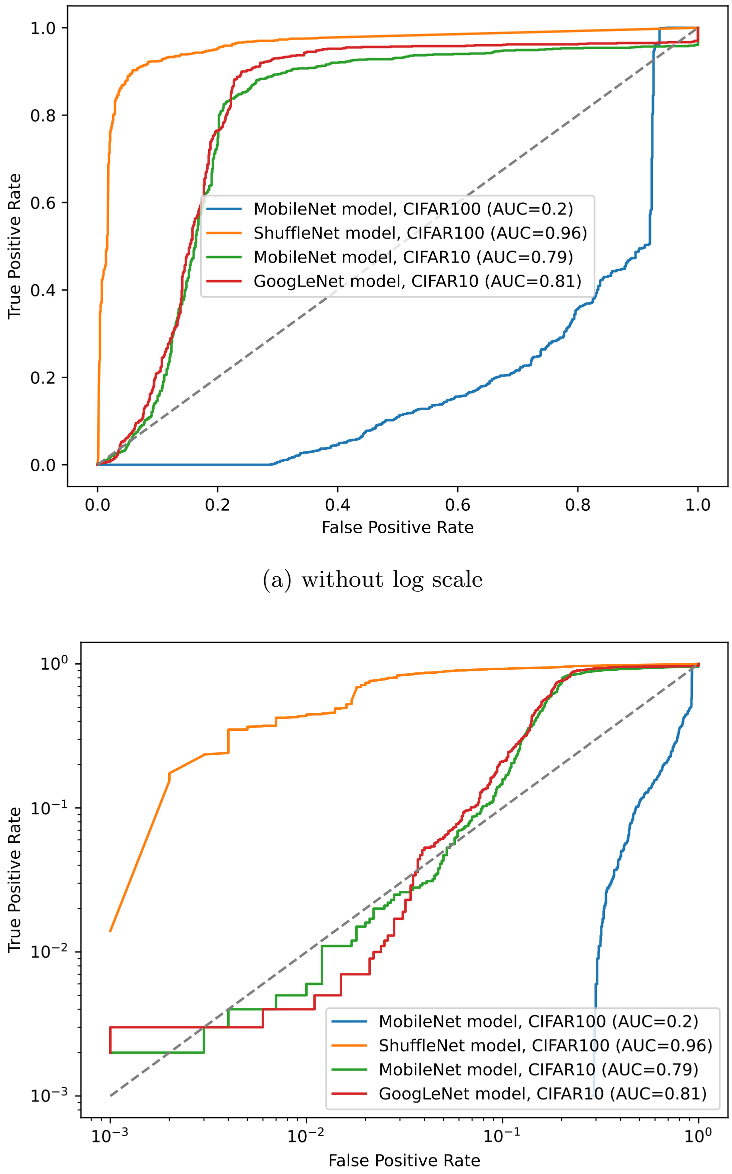
Figure 5.4: ROC curve without log scale and with log scale for the `Growing` attack with a control set of non-members.



(a) without log scale

(b) with log scale

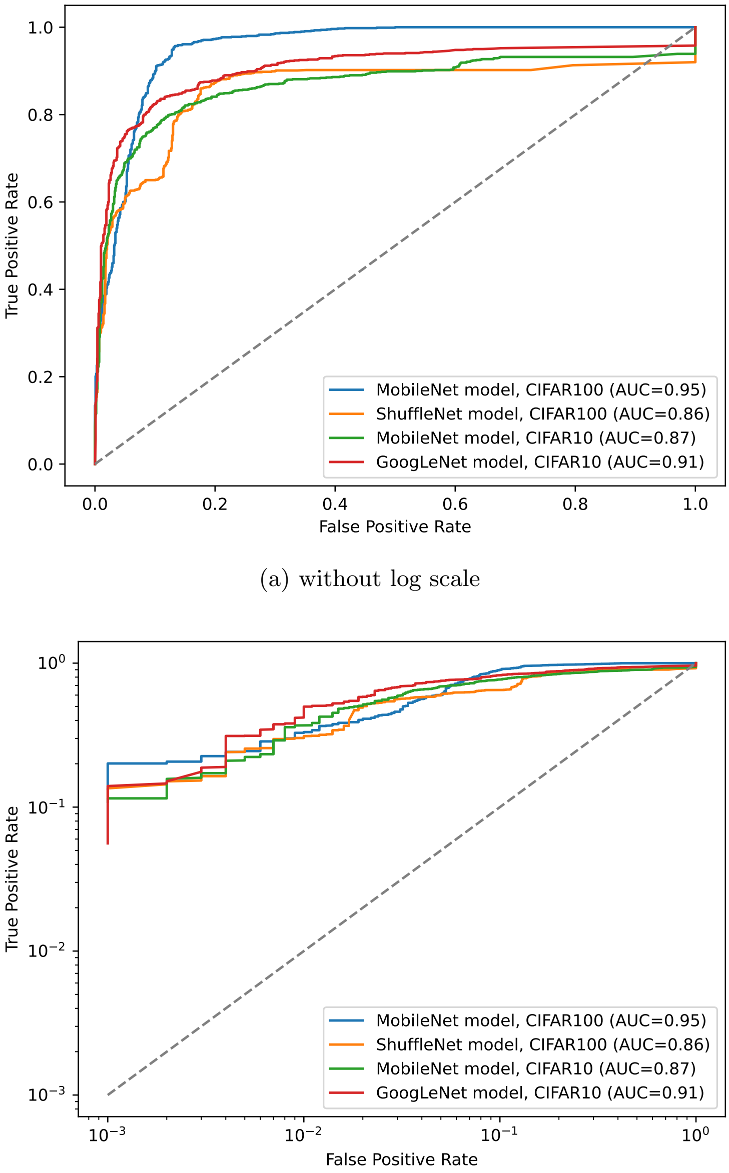
Figure 5.5: ROC curve without log scale and with log scale for the `Growing` attack with a control set of members.



(a) without log scale

(b) with log scale

Figure 5.6: ROC curve without log scale and with log scale for the `Replacement` attack with a control set of non-members.



(a) without log scale

(b) with log scale

Figure 5.7: ROC curve without log scale and with log scale for the `Replacement` attack with a control set of members.

6 Discussion

The following chapter discusses the results presented in the last chapter. It starts with the optimization of the regressors that are used in all novel attacks. Then, the failure of the p-value-based attack approaches is shortly reviewed, followed by the discussion of the methods to determine the thresholds. The `threshold-dependent MIA` and group-based MIAs (`Removal`, `Replacement`, `Growing`) are looked at and possible success and failure reasons are named. The chapter concludes with summarized answers to the research questions.

6.1 Optimized regressor

As all attacks rely on an efficient and accurate regressor, it was important to start the implementation by optimizing it.
In each case, a successful reduction of the amount of needed train data, and especially the number of members in their train data, was possible. While all regressors had an improved accuracy, the regressors trained on the CIFAR100 embeddings had a very high accuracy, around 10 % higher than the ones trained on the CIFAR10 embeddings. To avoid confusion, as explained in Section 4.2.1 in Chapter 4, when talking about the accuracy of the regressor the accuracy of a classifier on the regressor is described. The result falls in line with the in Chapter 2 described success factor of an MIA, that members and non-members are more easily distinguishable in datasets with many classes. To sum up, the optimized regressors fulfill the goal of reducing the amount of training data overall and especially members in the train data, while even outperforming the original regressors. Nevertheless, still 3000 non-members and 2000 members are needed for the training. Since in a real-life setting the fewer elements are needed in advance for an MIA, the better, it would be beneficial to find ways of optimizing the regressors even more in future research.

6.2 p-value-based attack approaches

In Chapter 3, it was theoretically explained and shown in experiments that the p-value is not applicable for novel MIAs based on the DIA. As a result, the p-valued-attack approaches are to be discarded because the initial experiments showed that they are not working properly. Instead, the `threshold-dependent MIA` and group-based attacks with the effect size are suitable alternatives.

6.3 Threshold determination methods

The presented methods to get the thresholds for the regressor output sums difference (for `threshold-dependent MIA`) and effect size difference (for `Removal`, `Replacement`, `Growing`) are using either the median or a clustering approach. When the distribution in the attack's test set is known, like in this thesis, the median outperforms the threshold through clustering. As argued in Chapter 4, the median should be often applicable and was therefore looked closer at in this thesis. The optimal found median thresholds vary considerably for the different target models, especially for the `threshold-dependent MIA`. It was found that most of the elements have scores at the two ends of the possible confidence scores, with only a few elements in between. This may explain why the median thresholds vary a lot between the two extremes, while it only slightly changes the actual classification results.

Over all attacks, the median-based threshold classifies members and non-members similarly well. This is expected because the median, as the middle of all values on a test set with half members and half non-members, should separate both groups equally well. However, for some sporadic attacks on specific target models presented in Chapter 5, the median did not work sufficiently well enough as the threshold. A possible explanation is that in these cases, the members and non-members had overlaps because their distances to the decision boundaries were sometimes close to each other. Another reason could be that because of the permanent changes to the elements of the set (e.g. in `Removal` the mixed set shrinks in size), the set changes in some cases so much in behavior that the specific fixed threshold does not reflect it well enough anymore.

This shows the need to perform further research on threshold determination methods.

6.4 Threshold-dependent MIA

The threshold-dependent MIA is a novel MIA that applies the DIA methods while dropping the hypothesis test. The results show that it works as an MIA in general, but better on the CIFAR100 than on the CIFAR10 embeddings. Similar to the better regressor performance on CIFAR100 embeddings, this follows the explained success factor in Chapter 2 that MIAs work better on datasets with many classes (like CIFAR100) due to the datasets' higher generalization errors than datasets with fewer classes (like CIFAR10): Because the regressor more accurately works on the CIFAR100 embeddings, the threshold-dependent MIA can more accurately divide between members' regressor outputs and non-members' regressor outputs.

6.5 Group-based MIAs with effect size

The group-based MIA with effect size (Removal, Replacement, Growing) is the second novel MIA approach introduced in this thesis. It applies the DIA methods and makes use of the effect size instead of the p-value. The results showed which ways of performing the attack work and which do not. The next subchapters name possible reasons for why some attacks failed and others succeeded.

6.5.1 Possible reasons for failures

The attacks Removal and Replacement with a non-member control set failed. Possible reasons could be a combination of two things:

Firstly, in contrast to the Growing attacks, both attacks start with a big mixed set, consisting of half members and half non-members. This makes it an ambiguous dataset, which might make it more likely that the effect size might not be always able to clearly differentiate whether a member or a non-member was removed/replaced. Even though the mixed set should become less diverse the more elements are removed/replaced, this step might not work properly because of the possible failure of the attack with an at the beginning ambiguous dataset: The attacks remove/replace the wrong elements (non-members instead of members) from the mixed set, this results in a smaller (or with more replaced elements) set, which nevertheless is ambiguous because it consists not of mostly non-members but still a mixture of elements and so the initial problem is continuing.

The second reason for the failures might be the non-member control set and its ambivalence. As mentioned in the DIA paper [28] and described in Chapter 2, members tend to have the maximum distance from the decision boundary. In contrast, non-members can be less clearly determinable, some might be closer to the decision boundary and others further away. A non-member control set might thus be more heterogeneous than a member control set and therefore disadvantageous to correctly determine the effect size, similar to the described possible problems with an ambiguous mixed set. An example that a non-member control set alone is not a reason for failure is that the `Growing` attack with a non-member control set works similarly well as with a member control set.

To conclude, the attacks might both be failing for the combination of the reasons: An ambiguous mixed set, combined with an ambiguous control set, might be too ambivalent in general for the attacks to succeed.

6.5.2 Possible reasons for success

The same possible reasons why `Removal` and `Replacement` with a non-member control set fail, might be why `Removal`, `Replacement`, `Growing` with a member control set, and `Growing` with a non-member control set succeed:

The attacks with a member control set might have the advantage, as mentioned above and in the DIA paper [28], that members have a usually maximized distance to the decision boundary. This makes a set full of members less ambiguous, which might be an advantage for successfully calculating the effect size. An indicator of what big of a difference a member control set can make in contrast to a non-member set is the `Removal` attacks: While the `Removal` attack fails to work with a non-member control set, the `Removal` attack with a member control set is the most successful of all attacks.

While the diverse mixed set might be one of the reasons why `Replacement` and `Removal` with a non-member set failed, the lack of ambivalence might be the reason why the other attacks succeeded.

In contrast to the `threshold-dependent MIA`, the group-based MIAs do not perform better on any specific target model type. This might be because the attacks are less straightforward, so their success is more independent of the target models' amount of classes and more dependent on how well the set and the control set are comparable via effect size.

6.6 Comparison of group-based and non-group-based MIAs

The most successful group-based attack (`Removal` with a member control set) has a TPR of 44.8 % at 0.1 % FPR (for the CIFAR100 ShuffleNet model). In comparison, the most successful non-group-based attack (`threshold-dependent MIA`, also on the CIFAR100 ShuffleNet model) has a TPR of 6.9 % at the same FPR. The best group-based attack is therefore 37.9 % better than the best non-group-based attack, making it a substantially more successful MIA. The superiority of the group-based MIA is as expected because of the potential of group-based approaches for determining the membership shown in this thesis (DIA [28] and BlindMIA [21], see Chapters 2, 3). The better performance validates the need for higher complexity in the group-based MIAs and shows that it should be focused on this attack type in future research.

6.7 Answers to research questions

Coming back to the research questions of this thesis, the answers can now be summarized as the following:

RQ1 *How can the findings and methods of the DIA be used for novel MIAs?*
As described in detail in Chapter 3, they can be used to perform a `threshold-dependent MIA` and group-based MIAs using the statistical effect size (`Removal`, `Replacement`, `Growing`). All attacks follow the DIA methods and first run the elements through the target model, create the embeddings with Blind Walk and use a regressor on the embeddings. The `threshold-dependent MIA` then separates the regressor outputs directly in predicted members and non-members, the group-based MIAs apply the effect size to predict the membership.

RQ2 *Is the DIA's group-based approach applicable for a novel group-based MIA?*
The approach until the hypothesis testing is applicable for an MIA, the hypothesis test with the p-value has to be dismissed and replaced by the effect size, as described in Chapter 3. The changed group-based approach (`Removal`, `Replacement`, `Growing`) is applicable for a novel attack.

RQ3 *How well do the novel MIA approaches perform?*
Four out of six group-based attacks and the `threshold-dependent MIA` succeed as MIAs.
`Removal` with a member control set performs overall the best and has the highest TPR at 0.1 % FPR of all attacks with 44.8 % (for the CIFAR100 ShuffleNet target model). `Removal` and `Replacement` with non-member control set fail to succeed.

The MIAs with the median as their specific threshold perform overall well, however, the attacks fail on certain target models for the `Replacement` with a non-member, member set, and `Growing` with a non-member control set. This is described in Chapter 5 and discussed in this chapter. In general, the group-based attacks are substantially more successful than the one working on individual points (`threshold-dependent MIA`), justifying the extra complexity needed for group comparisons.

7 Conclusion and Outlook

This thesis looked at the applicability of DIA methods for a novel MIA, especially for a group-based MIA.
It was argued that the p-value from the DIA was not applicable for the new MIA approaches and four new MIAs were presented: `threshold-dependent MIA`, `Removal`, `Growing` and `Replacement`, each with member and non-member control set. The latter ones used the effect size instead of the p-value, while the `threshold-dependent MIA` renounced the group-based approach altogether.
The attacks were evaluated by looking at the ROC curve with and without log scale, the AUC and the TPR at 0.1 % FPR.
`Removal` and `Replacement` with a non-member control set failed as attacks, as a possible reason the ambivalence of the sets was named. All other attacks succeeded, `Removal` with a member control set was found to be the most successful attack. It outperformed the non-group-based attack, highlighting the relevance of group-based MIAs.

7.1 Future research

In general, this thesis shows that future research should be focused on group-based MIAs (especially `Removal` with a member control set) as they outperformed the non-group-based attack (`threshold-dependent MIA`) in this thesis.

Specific to the presented attacks, it would be interesting to find ways to improve the regressor even more so that even less training data (which needs the creation of members and non-members) is needed.
Additionally, more methods to find the threshold for the attack, besides the median and clustering, should be looked at to find the optimal solution.
The attacks could further be evaluated on more datasets, models and different types of data besides image classification to get a deeper understanding of their applicability.
Another important next step to understand which value the attacks hold would be to compare them to other MIAs.

Besides the above-mentioned research in further developing the attacks, the described novel MIAs can be used for the evaluation of current and future defenses against MIAs, to protect the privacy of the machine learning model's sensitive training data.

Bibliography

[1] Charalampos Alexopoulos, Zoi Lachana, Aggeliki Androutsopoulou, Vasiliki Diamantopoulou, Yannis Charalabidis, and Michalis Avgerinos Loutsaris. “How Machine Learning is Changing e-Government”. In: *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance.* Melbourne VIC Australia: ACM, Apr. 2019, pp. 354–363. ISBN: 978-1-4503-6644-1. DOI: 10.1145/3326365.3326412. URL: https://dl.acm.org/doi/10.1145/3326365.3326412 (visited on 04/10/2022).

[2] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* en. Google-Books-ID: bX3TBwAAQBAJ. Springer Science & Business Media, June 2011. ISBN: 978-1-4419-9096-9. DOI: 10.1007/978-1-4419-9096-9.

[3] Rohan Bhardwaj, Ankita R. Nambiar, and Debojyoti Dutta. “A Study of Machine Learning in Healthcare”. In: *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC).* Vol. 2. ISSN: 0730-3157. July 2017, pp. 236–241. DOI: 10.1109/COMPSAC.2017.164.

[4] Stephan B. Bruns and John P. A. Ioannidis. “p-Curve and p-Hacking in Observational Research”. en. In: *PLOS ONE* 11.2 (Feb. 2016). Ed. by Daniele Marinazzo, e0149144. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0149144. URL: https://dx.plos.org/10.1371/journal.pone.0149144 (visited on 07/20/2022).

[5] Ben Van Calster, Ewout W Steyerberg, Gary S Collins, and Tim Smits. “Consequences of relying on statistical significance: some illustrations”. en. In: *European journal of clinical investigation* 48(5) (2018), p. 21. DOI: 10.1111/eci.12912.

[6] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. “Membership Inference Attacks From First Principles”. In: *2022 IEEE Symposium on Security and Privacy (SP)* (2022). arXiv: 2112.03570, pp. 1897–1914. URL: http://arxiv.org/abs/2112.03570 (visited on 03/19/2022).

[7] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. “GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models”. en. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security.* Virtual Event USA: ACM, Oct. 2020, pp. 343–362. ISBN:

978-1-4503-7089-9. DOI: 10.1145/3372297.3417238. URL: https://dl.acm.org/doi/10.1145/3372297.3417238 (visited on 09/23/2022).

[8] David Burton and John Coleman. "Quasi-Cauchy Sequences". en. In: *The American Mathematical Monthly* 117.4 (2010), p. 328. ISSN: 00029890. DOI: 10.4169/000298910x480793. URL: https://www.tandfonline.com/doi/full/10.4169/000298910X480793 (visited on 10/24/2022).

[9] Eugene Demidenko. "The P -value you can't buy". In: *The American Statistician* 70 (July 2015), pp. 33–38. DOI: 10.1080/00031305.2015.1069760.

[10] Paul D. Ellis. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results.* en. 1st ed. Cambridge University Press, July 2010, pp. 7–11, 26. ISBN: 978-0-521-19423-5 978-0-521-14246-5 978-0-511-76167-6. DOI: 10.1017/CBO9780511761676. URL: https://www.cambridge.org/core/product/identifier/9780511761676/type/book (visited on 08/05/2022).

[11] Tom Fawcett. "An introduction to ROC analysis". en. In: *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S016786550500303X (visited on 09/25/2022).

[12] Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. *Adversarial Attacks Against Medical Deep Learning Systems.* arXiv preprint arXiv: 1804.05296. 2018. URL: http://arxiv.org/abs/1804.05296 (visited on 03/15/2022).

[13] Andrew Gelman and Hal Stern. "The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant". en. In: *The American Statistician* 60.4 (Nov. 2006), pp. 328–331. ISSN: 0003-1305, 1537-2731. DOI: 10.1198/000313006X152649. URL: http://www.tandfonline.com/doi/abs/10.1198/000313006X152649 (visited on 07/07/2022).

[14] Benyamin Ghojogh and Mark Crowley. *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial.* arXiv preprint arXiv:1905.12787. May 2019. DOI: 10.48550/ARXIV.1905.12787. URL: http://arxiv.org/abs/1905.12787 (visited on 06/17/2022).

[15] Steven Goodman. "A Dirty Dozen: Twelve P-Value Misconceptions". en. In: *Seminars in Hematology* 45.3 (July 2008), pp. 135–140. ISSN: 00371963. DOI: 10.1053/j.seminhematol.2008.04.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0037196308000620 (visited on 07/05/2022).

[16] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". en. In: *European Journal of Epidemiology* 31.4 (Apr. 2016), pp. 337–350. ISSN: 0393-2990, 1573-7284. DOI: 10.1007/s10654-016-0149-3. URL: http://link.springer.com/10.1007/s10654-016-0149-3 (visited on 07/20/2022).

[17] Allan Grønlund, Kasper Green Larsen, Alexander Mathiasen, Jesper Sindahl Nielsen, Stefan Schneider, and Mingzhou Song. *Fast Exact k-Means, k-Medians and Bregman Divergence Clustering in 1D*. en. arXiv preprint arXiv:1701.07204 [cs]. Apr. 2018. URL: http://arxiv.org/abs/1701.07204 (visited on 09/22/2022).

[18] Lewis G. Halsey. "The reign of the p -value is over: what alternative analyses could we employ to fill the power vacuum?" en. In: *Biology Letters* 15.5 (May 2019), p. 20190174. ISSN: 1744-9561, 1744-957X. DOI: 10.1098/rsbl.2019.0174. URL: https://royalsocietypublishing.org/doi/10.1098/rsbl.2019.0174 (visited on 07/08/2022).

[19] Lewis G. Halsey, Douglas Curran-Everett, Sarah L. Vowler, and Gordon B. Drummond. "The fickle P value generates irreproducible results". en. In: *Nature Methods* 12.3 (Mar. 2015). Number: 3 Publisher: Nature Publishing Group, pp. 179–185. ISSN: 1548-7105. DOI: 10.1038/nmeth.3288. URL: https://www.nature.com/articles/nmeth.3288 (visited on 07/20/2022).

[20] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. *Mobilenets: Efficient convolutional neural networks for mobile vision applications*. arXiv preprint arXiv:1704.04861. 2017.

[21] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. "Practical Blind Membership Inference Attack via Differential Comparisons". In: *Proceedings 2021 Network and Distributed System Security Symposium*. 2021. ISBN: 978-1-891562-66-2. DOI: 10.14722/ndss.2021.24293. URL: https://www.ndss-symposium.org/wp-content/uploads/ndss2021_5C-2_24293_paper.pdf (visited on 03/14/2022).

[22] Paul Irolla and Gregory Chatel. "Demystifying the Membership Inference Attack". en. In: *2019 12th CMI Conference on Cybersecurity and Privacy (CMI)*. Copenhagen, Denmark: IEEE, Nov. 2019, pp. 1–7. ISBN: 978-1-72812-856-6. DOI: 10.1109/CMI48017.2019.8962136. URL: https://ieeexplore.ieee.org/document/8962136/ (visited on 03/14/2022).

[23] Daniel Jakubovitz, Raja Giryes, and Miguel RD Rodrigues. "Generalization error in deep learning". In: *Compressed sensing and its applications*. Springer, 2019, pp. 153–193.

[24] P. N. Jani. *Business Statistics: Theory and Applications*. ar. Google-Books-ID: fH5EBQAAQBAJ. PHI Learning Pvt. Ltd., Sept. 2014. ISBN: 978-81-203-4985-8.

[25] Dong Kyu Lee. "Alternatives to P value: confidence interval and effect size". In: *Korean Journal of Anesthesiology* 69.6 (Oct. 2016). Publisher: The Korean Society of Anesthesiologists, pp. 555–562. DOI: 10.4097/kjae.2016.69.6.555. URL: https://synapse.koreamed.org/articles/1156285 (visited on 08/05/2022).

[26] Johnson Ching-Hong Li. “Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data”. In: *Behavior research methods* 48.4 (2016), pp. 1560–1574.

[27] Mingfeng Lin, Henry Lucas, and Galit Shmueli. “Too Big to Fail: Large Samples and the p-Value Problem”. In: *Information Systems Research* 24 (Dec. 2013), pp. 906–917. DOI: 10.1287/isre.2013.0480.

[28] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. “Dataset Inference: Ownership Resolution in Machine Learning”. In: *International Conference on Learning Representations* (Apr. 2021). arXiv: 2104.10706. URL: http://arxiv.org/abs/2104.10706 (visited on 03/14/2022).

[29] Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha. “Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare”. In: *IEEE Journal of Biomedical and Health Informatics* 19.6 (Nov. 2015), pp. 1893–1905. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2014.2344095. URL: http://ieeexplore.ieee.org/document/6868201/ (visited on 03/15/2022).

[30] Ivens Portugal, Paulo Alencar, and Donald Cowan. “The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review”. In: *Expert Systems with Applications* 97 (2018). arXiv: 1511.05263, pp. 205–227. URL: http://arxiv.org/abs/1511.05263 (visited on 04/10/2022).

[31] Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. “Differential privacy defenses and sampling attacks for membership inference”. In: *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security.* 2021, pp. 193–202.

[32] 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”. en. In: *Proceedings 2019 Network and Distributed System Security Symposium.* San Diego, CA: Internet Society, 2019. ISBN: 978-1-891562-55-6. DOI: 10.14722/ndss.2019.23119. URL: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03A-1_Salem_paper.pdf (visited on 03/14/2022).

[33] “Mean Squared Error”. In: *Encyclopedia of Machine Learning.* Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 653–653. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_528. URL: https://doi.org/10.1007/978-0-387-30164-8_528.

[34] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. “Membership Inference Attacks are Easier on Difficult Problems”. en. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* Montreal, QC, Canada: IEEE, Oct. 2021, pp. 14800–14809. ISBN: 978-1-66542-812-5. DOI: 10.1109/ICCV48922.2021.01455. URL: https://ieeexplore.ieee.org/document/9711338/ (visited on 06/13/2022).

[35] Muhammad A Shah, Joseph Szurley, Markus Mueller, Athanasios Mouchtaris, and Jasha Droppo. "Evaluating the Vulnerability of End-to-End Automatic Speech Recognition Models To Membership Inference Attacks". en. In: *Interspeech*. 2021, pp. 891–895.

[36] Virat Shejwalkar, Huseyin A Inan, Amir Houmansadr, and Robert Sim. "Membership Inference Attacks Against NLP Classification Models". en. In: *NeurIPS 2021 Workshop Privacy in Machine Learning*. 2021.

[37] Yun Shen, Xinlei He, Yufei Han, and Yang Zhang. "Model stealing attacks against inductive graph neural networks". In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 1175–1192.

[38] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks against Machine Learning Models". In: *2017 IEEE symposium on security and privacy (SP)*. arXiv: 1610.05820. IEEE. Mar. 2017, pp. 3–18. URL: http://arxiv.org/abs/1610.05820 (visited on 03/19/2022).

[39] Gail M. Sullivan and Richard Feinn. "Using Effect Size—or Why the P Value Is Not Enough". In: *Journal of Graduate Medical Education* 4.3 (Sept. 2012), pp. 279–282. ISSN: 1949-8349. DOI: 10.4300/JGME-D-12-00156.1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/ (visited on 07/08/2022).

[40] Roberto Szechtman. "Chapter 10 A Hilbert Space Approach to Variance Reduction". en. In: *Handbooks in Operations Research and Management Science*. Vol. 13. Elsevier, 2006, pp. 259–289. ISBN: 978-0-444-51428-8. DOI: 10.1016/S0927-0507(06)13010-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S0927050706130108 (visited on 07/08/2022).

[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[42] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. *Towards Demystifying Membership Inference Attacks*. en. arXiv preprint arXiv:1807.09173. Feb. 2019. URL: http://arxiv.org/abs/1807.09173 (visited on 03/19/2022).

[43] Samuel Yeom, Irene Giacomelli, Alan Menaged, Matt Fredrikson, and Somesh Jha. "Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning". In: *Journal of Computer Security* 28.1 (Feb. 2020), pp. 35–70. ISSN: 18758924, 0926227X. DOI: 10.3233/JCS-191362. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JCS-191362 (visited on 03/13/2022).

[44] Xiaoyong Yuan, Leah Ding, Lan Zhang, Xiaolin Li, and Dapeng Oliver Wu. "ES attack: Model stealing against deep neural networks without data hurdles". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* (2022).

[45] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. “Shufflenet: An extremely efficient convolutional neural network for mobile devices”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018, pp. 6848–6856.

Appendix

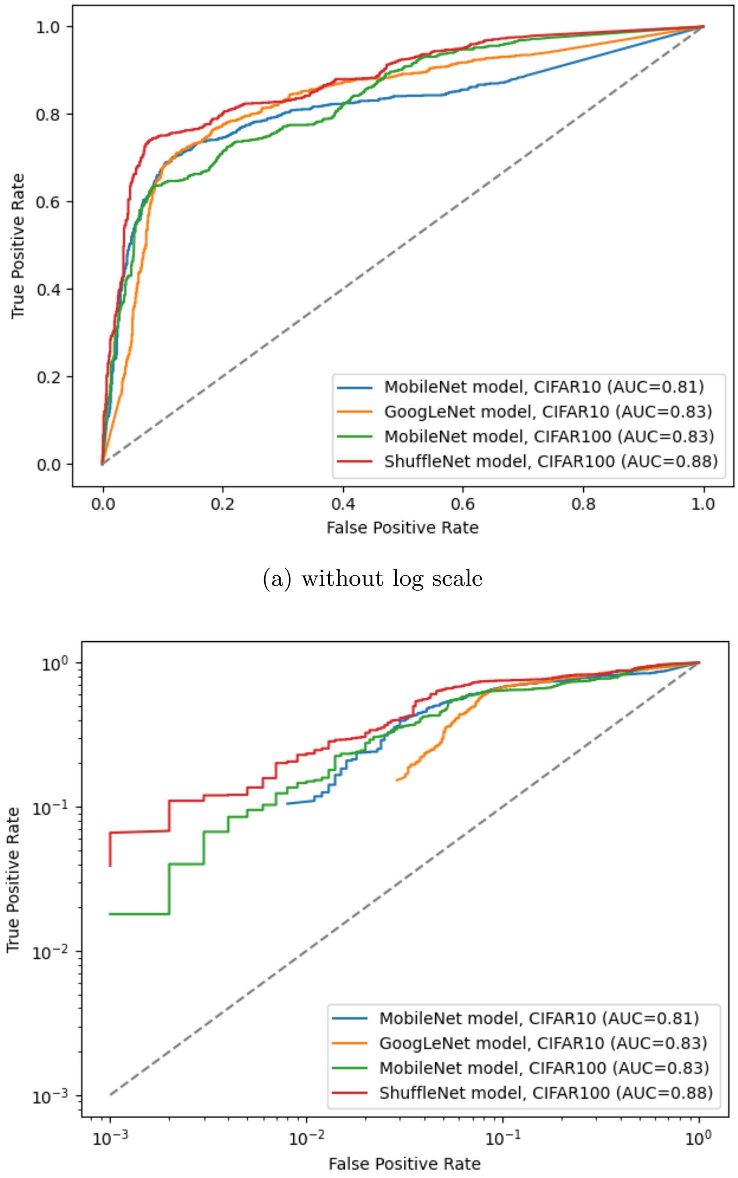
7.2 Appendix

7.2.1 Threshold-dependent MIA, threshold through clustering

The found threshold through clustering for CIFAR10 MobileNet is 0.0268, for CIFAR10 GoogLeNet 0.0068, 0.0615 for CIFAR100 MobileNet and -0.0363 for CIFAR100 ShuffleNet.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Threshold	**Threshold through clustering**			
TPR at 0.1 % FPR	Not found, 0.1 % FPR does not exist. TPRs at closest existing FPRs: At 0.8 % FPR: 10.5% At 0 % FPR: 0%	Not found, 0.1 % FPR does not exist. TPRs at closest existing FPRs: At 2.9 % FPR: 15.3% At 0 % FPR: 0%	1.80%	3.90%
AUC	0.81	0.83	0.83	0.88
Accuracy	77.45%	79.00%	70.95%	82.80%
F1-score	73.70%	76.45%	73.84%	81.32%
Amount correctly classified members (out of 1000 members)	632	682	820	749
Amount correctly classified non-members (out of 1000 non-members)	917	898	599	907

Table 7.1: Performance of the `threshold-dependent MIA`. When 0.1 % FPR does not exist, closest FPRs are mentioned.



(a) without log scale

(b) with log scale

Figure 7.1: ROC curve without log scale and with log scale for the `threshold-dependent MIA`.

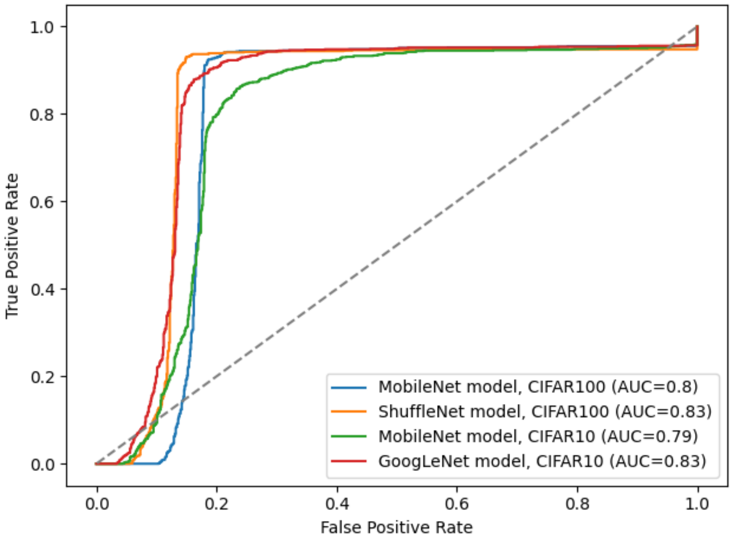
7.2.2 Group-based MIA with effect size, threshold through clustering

Model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Datatset	**CIFAR10**		**CIFAR100**	
Threshold type	**Threshold through clustering**			
Removal (non-member control set)	-4.77E-05	-5.36E-05	-9.03E-05	-7.36E-05
Removal (member control set)	7.80E-05	5.74E-05	-1.38E-05	-2.96E-06
Growing (non-member control set)	-0.003194736636272	-0.002742686675566	0.002217660043317	-0.001816807305138
Growing (member control set)	3.83E-04	3.32E-04	1.85E-03	9.80E-04
Replacement (non-member control set)	-3.12E-05	-1.80E-05	-9.03E-05	-7.05E-05
Replacement (member control set)	6.24E-05	5.97E-05	-1.49E-05	1.68E-05

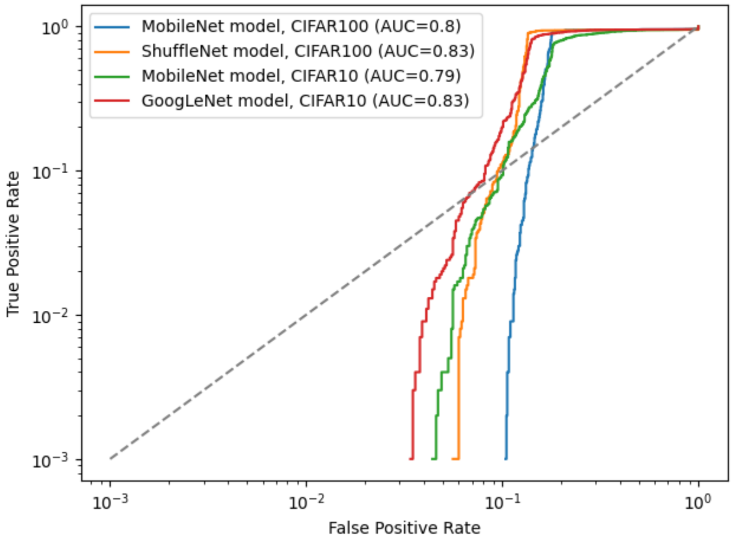
Table 7.2: Optimal thresholds for group-based attacks.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Non-member**			
Threshold	**Threshold through clustering**			
TPR at 0.1 % FPR	0.00%	0.00%	0.00%	0.00%
AUC	0.79	0.83	0.8	0.83
Accuracy	79.35%	85.15%	85.85%	88.80%
F1-score	81.25%	86.14%	86.94%	89.32%
Amount correctly classified members (out of 1000 members)	895	923	942	937
Amount correctly classified non-members (out of 1000 non-members)	692	780	775	839

Table 7.3: Performance of the `Removal` attack with a control set of non-members.

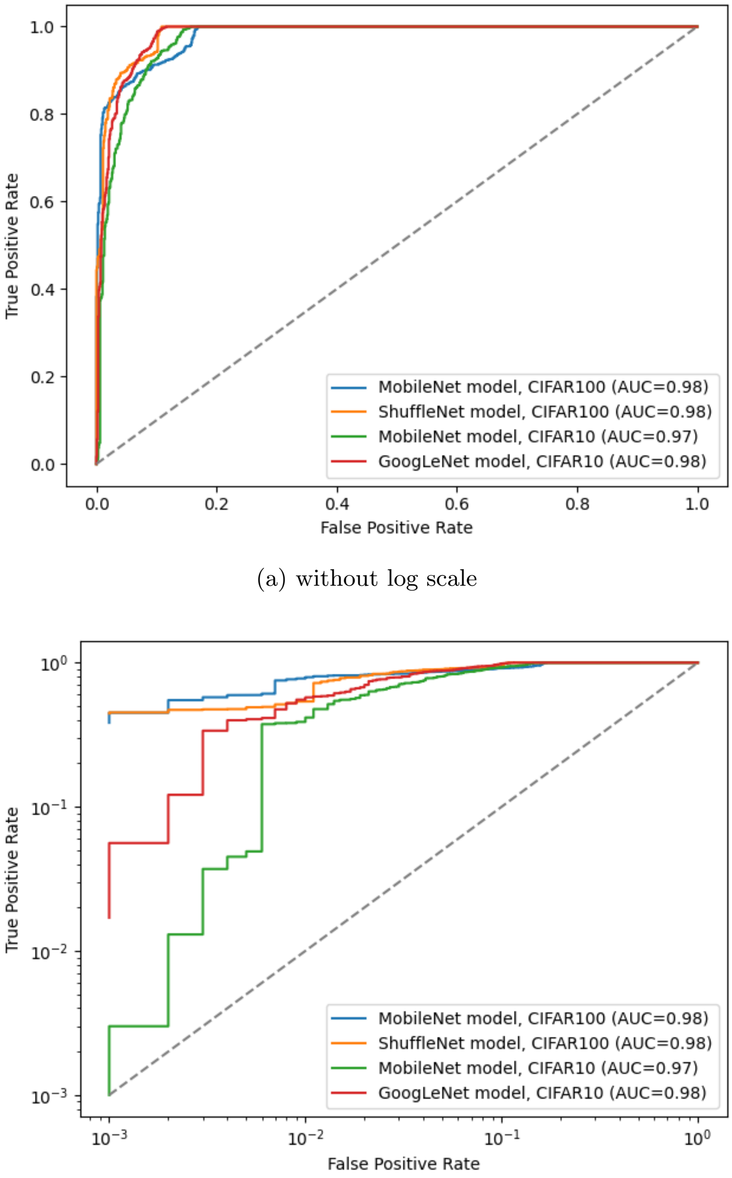


(a) without log scale



(b) with log scale

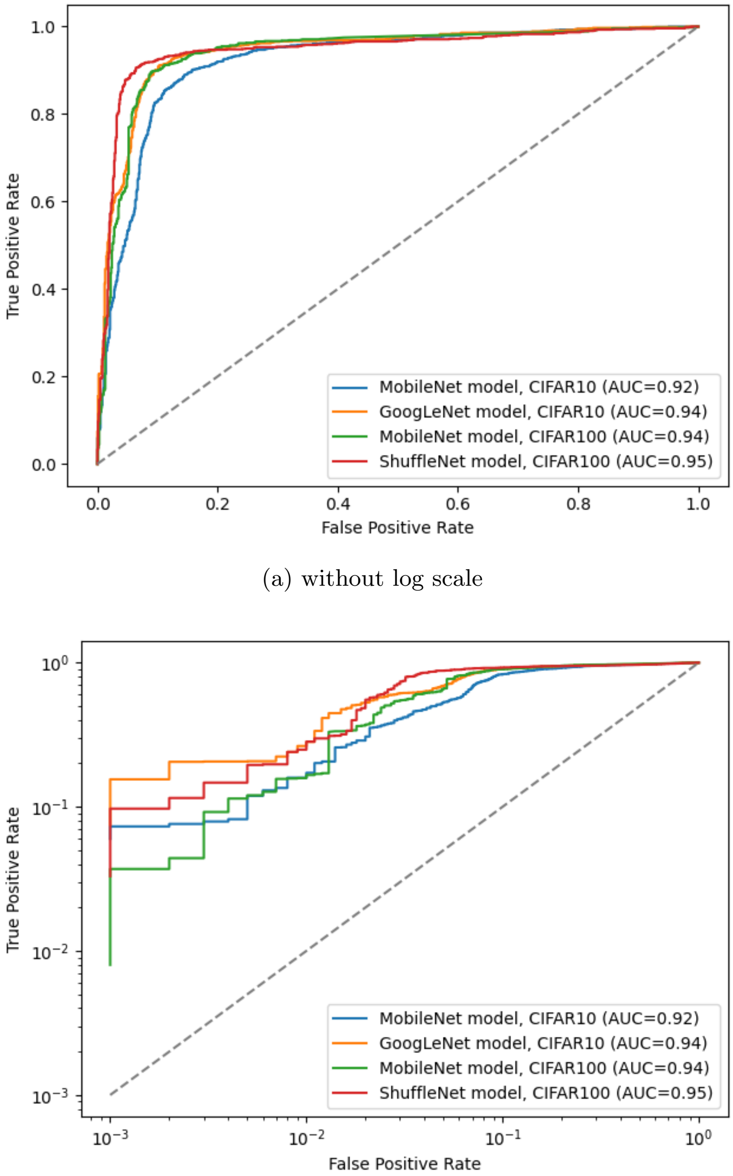
Figure 7.2: ROC curve without log scale and with log scale for the `Removal` attack with a control set of non-members.



(a) without log scale

(b) with log scale

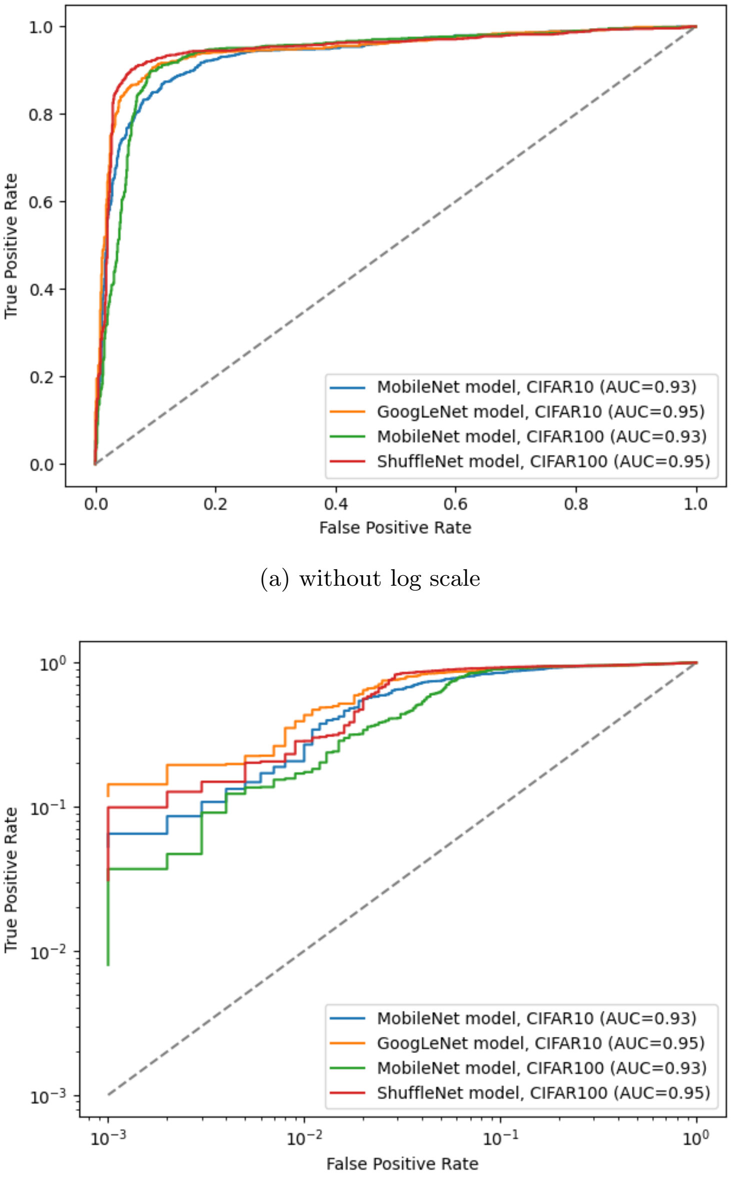
Figure 7.3: ROC curve without log scale and with log scale for the `Removal` attack with a control set of members.



(a) without log scale

(b) with log scale

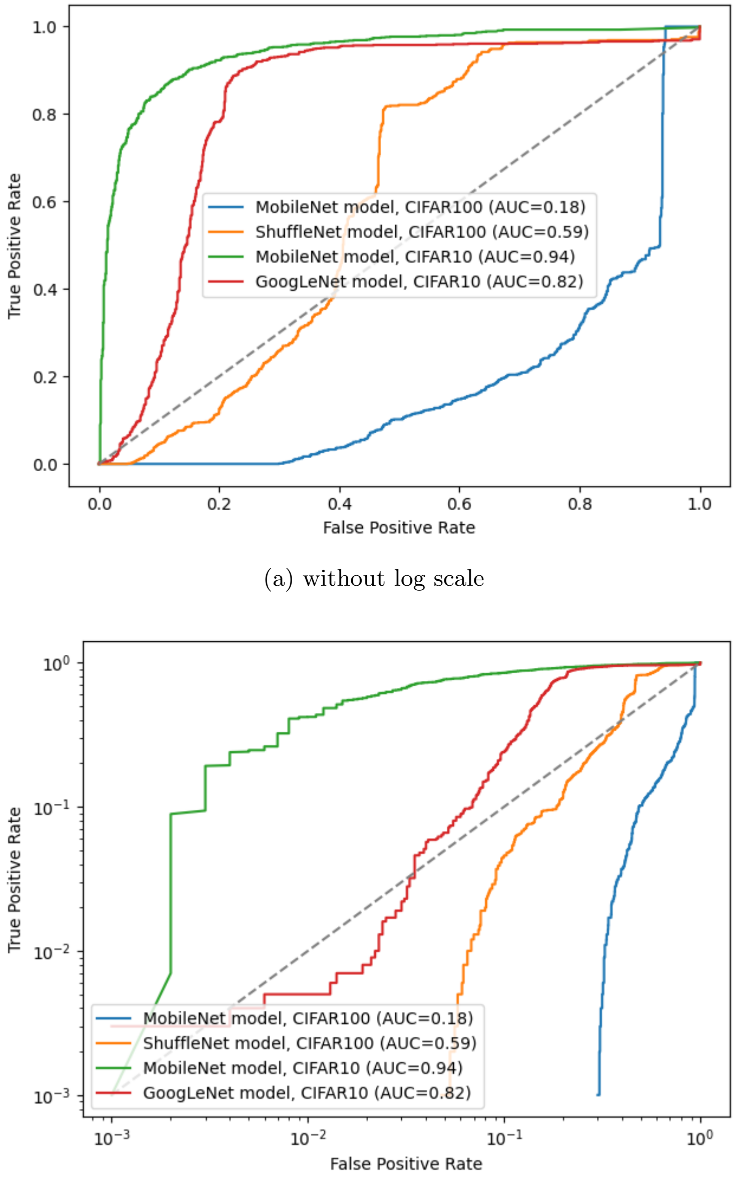
Figure 7.4: ROC curve without log scale and with log scale for the `Growing` attack with a control set of non-members.



(a) without log scale

(b) with log scale

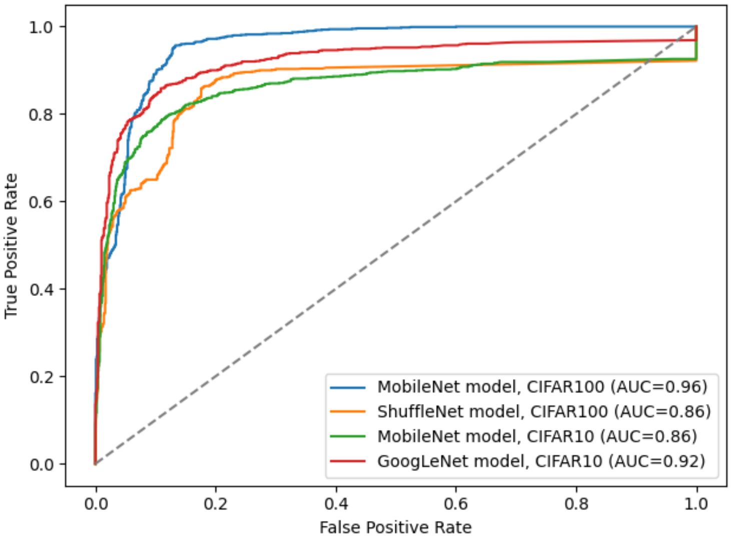
Figure 7.5: ROC curve without log scale and with log scale for the `Growing` attack with a control set of members.



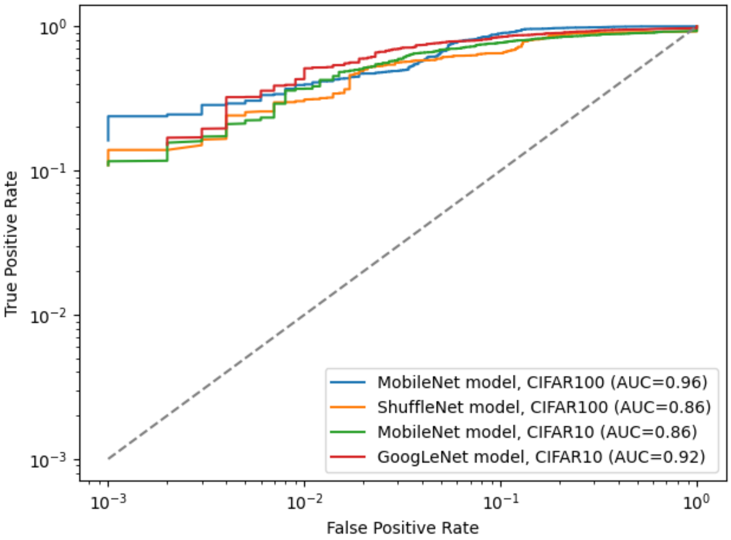
(a) without log scale

(b) with log scale

Figure 7.6: ROC curve without log scale and with log scale for the `Replacement` attack with a control set of non-members.



(a) without log scale



(b) with log scale

Figure 7.7: ROC curve without log scale and with log scale for the `Replacement` attack with a control set of members.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Member**			
Threshold	**Threshold through clustering**			
TPR at 0.1 % FPR	0.10%	1.70%	38.30%	44.40%
AUC	0.97	0.98	0.98	0.98
Accuracy	86.80%	90.40%	90.30%	92.75%
F1-score	85.53%	89.81%	90.46%	92.63%
Amount correctly classified members (out of 1000 members)	780	846	920	920
Amount correctly classified non-members (out of 1000 non-members)	956	962	886	886

Table 7.4: Performance of the `Removal` attack with a control set of members.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Non-member**			
Threshold	**Threshold through clustering**			
TPR at 0.1 % FPR	6.00%	7.50%	0.80%	3.30%
AUC	0.92	0.94	0.94	0.95
Accuracy	65.20%	74.95%	59.10%	91.90%
F1-score	48.29%	67.40%	70.74%	91.88%
Amount correctly classified members (out of 1000 members)	325	518	989	917
Amount correctly classified non-members (out of 1000 non-members)	979	981	193	921

Table 7.5: Performance of the `Growing` attack with a control set of non-members.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Member**			
Threshold	**Threshold through clustering**			
TPR at 0.1 % FPR	5.30%	11.90%	0.80%	3.10%
AUC	0.93	0.95	0.93	0.95
Accuracy	86.05%	89.50%	88.95%	91.70%
F1-score	84.89%	88.77%	89.44%	91.72%
Amount correctly classified members (out of 1000 members)	784	830	936	920
Amount correctly classified non-members (out of 1000 non-members)	937	960	843	914

Table 7.6: Performance of the `Growing` attack with a control set of members.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Non-member**			
Threshold	**Threshold through clustering**			
TPR at 0.1 % FPR	0.10%	0.30%	0.00%	0.00%
AUC	0.94	0.82	0.18	0.59
Accuracy	50.05%	82.85%	52.35%	64.75%
F1-score	0.19%	83.63%	67.50%	72.39%
Amount correctly classified members (out of 1000 members)	1	876	990	924
Amount correctly classified non-members (out of 1000 non-members)	1000	781	57	371

Table 7.7: Performance of the `Replacement` attack with a control set of non-members.

Target model	**MobileNet**	**GoogLeNet**	**MobileNet**	**ShuffleNet**
Dataset	**CIFAR10**		**CIFAR100**	
Control set	**Member**			
Threshold	**Threshold through clustering**			
TPR at 0.1 % FPR	10.90%	15.10%	16.20%	10.90%
AUC	0.86	0.92	0.96	0.86
Accuracy	48.75%	48.85%	90.45%	47.65%
F1-score	65.54%	65.63%	90.74%	64.54%
Amount correctly classified members (out of 1000 members)	975	977	936	953
Amount correctly classified non-members (out of 1000 non-members)	0	0	873	0

Table 7.8: Performance of the `Replacement` attack with a control set of members.