Understanding Data Importance in Machine Learning Attacks: Does Valuable Data Pose Greater Harm?

Rui Wen
CISPA Helmholtz Center for
Information Security
rui.wen@cispa.de

Michael Backes
CISPA Helmholtz Center for
Information Security
director@cispa.de

Yang Zhang*
CISPA Helmholtz Center for
Information Security
zhang@cispa.de

***Abstract*—Machine learning has revolutionized numerous domains, playing a crucial role in driving advancements and enabling data-centric processes. The significance of data in training models and shaping their performance cannot be overstated. Recent research has highlighted the heterogeneous impact of individual data samples, particularly the presence of valuable data that significantly contributes to the utility and effectiveness of machine learning models. However, a critical question remains unanswered: are these valuable data samples more vulnerable to machine learning attacks? In this work, we investigate the relationship between data importance and machine learning attacks by analyzing five distinct attack types. Our findings reveal notable insights. For example, we observe that high importance data samples exhibit increased vulnerability in certain attacks, such as membership inference and model stealing. By analyzing the linkage between membership inference vulnerability and data importance, we demonstrate that sample characteristics can be integrated into membership metrics by introducing sample-specific criteria, therefore enhancing the membership inference performance. These findings emphasize the urgent need for innovative defense mechanisms that strike a balance between maximizing utility and safeguarding valuable data against potential exploitation.**

I. Introduction

Machine learning has emerged as an indispensable tool across numerous domains, revolutionizing industries and empowering data-driven decision-making processes. Central to the essence of machine learning is the pivotal role played by data, serving as the bedrock for training models and exerting a profound influence on their performance and predictive accuracy. Concurrently, the crucial role of data in model training also exposes it as a noteworthy source of vulnerabilities.

Recent research has shed light on the heterogeneous impact of individual data samples, highlighting the presence of certain data that exhibit a heightened influence on the utility and overall effectiveness of machine learning models [24], [16], [28], [27], [25]. Understanding this variability is important for two main reasons. First, knowing how individual data samples affect model performance is key to improving machine learning explainability, offering new insights into model behavior, and enhancing interpretability [27], [47], [37], [46]. Second, this knowledge can guide data trading practices, where the importance of data is a significant factor [4], [16].

However, the influence of such diverse data on model leakage and security remains largely unexplored. Existing research predominantly concentrates on the models themselves; for example, studies [50], [35] suggest that overfitted models are more prone to membership inference attacks. Nevertheless, even within the same model, distinct data samples exhibit varying vulnerabilities to attacks. This prompts a crucial question: do these valuable data samples also exhibit an increased vulnerability to a spectrum of machine learning attacks? Understanding the differential vulnerability of data samples has significant practical implications. In medical diagnostics, for example, patient records with rare but highly indicative symptoms are considered high importance samples. Assessing whether these records are more prone to attacks is crucial, as breaches could lead to discrimination, higher insurance premiums, or other serious consequences for individuals.

In this paper, our focus lies in investigating the relationship between data importance and machine learning attacks. Our primary objective is to thoroughly investigate whether valuable data samples, which contribute significantly to the utility of machine learning models, are exposed to an elevated risk of exploitation by malicious actors.

To achieve our objectives, we focus on five distinct types of attacks, encompassing both training-time and testing-time attacks. The training-time attack we consider is the backdoor attack [17], [32], [49], while the testing-time attacks consist of membership inference attack [34], [55], [50], [12], [31], model stealing attack [63], [62], [52], attribute inference attack [39], [57], and data reconstruction attack [75], [72], [14]. For each of these attacks, we thoroughly analyze the behavior and impact on both high importance and low importance data samples, aiming to uncover any discernible differences.

Main Findings: Our research has yielded significant findings that shed light on the heightened vulnerability of valuable data samples to privacy attacks. Specifically, our key findings are as follows:

- Membership Inference Attack: High importance data samples exhibit a higher vulnerability compared to low importance samples, particularly in the low false-positive rate region. For instance, in the CIFAR10

*Corresponding author.

Network and Distributed System Security (NDSS) Symposium 2025
23 - 28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.23331
www.ndss-symposium.org

dataset, at a false positive rate (FPR) of 1%, the true positive rate (TPR) of high importance data is $10.2\times$ greater than that of low importance samples.

- Privacy Onion Effect: The concept of the privacy onion effect [9] can be extended to the distribution of data importance. Specifically, previously considered unimportant samples gain significance when the dataset removes the important samples.
- Model Stealing Attack: High importance samples demonstrate greater efficiency in stealing models when the target model is trained on the same distribution as the query distribution. However, we empirically demonstrate that the importance does not transfer between different tasks.
- Backdoor Attack: Poisoning high importance data enhances the efficiency of the poisoning process, particularly when the size of the poison is small. On the other hand, the influence on clean accuracy does not yield a definitive conclusion, poisoning either type of data has a limited impact on clean accuracy.
- Attribute Inference and Data Reconstruction Attacks: We observe no significant distinction between high and low importance data in these attacks.

Our research provides empirical evidence establishing a correlation between data importance and vulnerabilities across diverse attack scenarios. This introduces a novel perspective for analyzing sample-specific vulnerabilities, enriching our understanding of the security implications within the realm of machine learning.

Beyond theoretical insights, our study showcases practical applications of these findings, illustrating how they can be utilized to devise more potent attacks. On one hand, these findings can be utilized in a *passive* manner. For example, we empirically demonstrate that membership inference attacks can be improved by introducing sample-specific criteria based on sample importance. Additionally, adjusting the poisoning strategy according to sample importance proves to enhance the efficacy of backdoor attacks, particularly with a reduced poisoning rate.

More interestingly, we can *actively* modify samples to alter their importance, which subsequently impacts both attack and defense performance. For example, recognizing that high importance samples are more vulnerable to membership inference attacks, attackers could increase the importance of targeted samples to heighten their vulnerability. This approach follows exactly the same idea as the attack accepted at CCS'22 [61], effectively demonstrating how we can "reinvent" state-of-the-art attacks guided by findings in our work.

In summary, our work represents a pioneering step in systematically understanding the vulnerabilities of the machine learning ecosystem through the lens of data. These findings serve as a resounding call to action, urging researchers and practitioners to develop innovative defenses that strike a delicate balance between maximizing utility and safeguarding valuable data against malicious exploitation.

II. Background

A. Machine Learning Models

Machine learning algorithms aim to construct models that effectively predict outputs based on given inputs. These models are typically represented by a parameterized function denoted as $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ represents the input space and $\mathcal{Y}$ represents the output space encompassing all possible predictions. The process of determining optimal parameter values θ involves minimizing an objective function using gradient descent. Specifically, the objective is to minimize the classification loss

$$\mathop{\mathbb{E}}_{(x,y)}[\mathcal{L}(f_\theta(x), y)]$$

where $(x, y) \in \mathcal{X} \times \mathcal{Y}$ denotes samples from the training dataset used to train the target model. This optimization process guides the model towards achieving optimal performance by iteratively adjusting the parameters.

B. Data Importance

The investigation of individual training sample importance in machine learning (ML) is a fundamental and intricate problem with broad implications, especially in data valuation. Understanding the importance of a single training sample within a learning task profoundly impacts data assessment, allocation of resources, and the quality of ML models.

Leave-one-out (LOO) method has long been regarded as an intuitive approach for assessing the importance of data samples. Formally, let D and D_{val} represent the training set and the validation set, and $\mathcal{A}$ denote the learning algorithm. $U_{\mathcal{A},D_{\text{val}}}$ denotes the validation accuracy of the model trained on D using $\mathcal{A}$. The importance of a target sample z can be quantified as the difference in utility before and after incorporating the target sample into the training set, expressed as:

$$v_{loo}(z) \propto U_{\mathcal{A},D_{val}}(D) - U_{\mathcal{A},D_{val}}(D \backslash \{z\})$$

Nevertheless, evaluating the importance of all N samples in the training set necessitates retraining the model N times, resulting in computational heaviness. To address this limitation, Koh and Liang [27] proposed influence functions as an approximation method, significantly reducing the computational cost from $O(Np^2 + p^3)$ to $O(Np)$, where p represents the number of model parameters.

Despite the effectiveness of LOO, Ghorbani and Zou [16] have raised concerns about its ability to capture complex interactions between subsets of data. They argue that the Shapley value provides a more comprehensive framework for measuring data importance. The Shapley value, originally proposed by Shapley [54], assigns an importance value to each sample z in the training set using the following formulation:

$$\nu_{\text{shap}}(z) \propto \frac{1}{N} \sum_{S \subseteq D \backslash \{z\}} \frac{1}{\binom{N-1}{|S|}} \left[U_{\mathcal{A},D_{val}}(S \cup \{z\}) - U_{\mathcal{A},D_{val}}(S)\right]$$

To simplify the interpretation of the Shapley value assigned to each sample, one can conceptualize it as the contribution to accuracy in typical scenarios.

For instance, in a hypothetical scenario with 100 samples and a model achieving 90% accuracy, a valuable sample may contribute 2% accuracy, while a less valuable sample may only contribute 0.1%. Consequently, the importance value assigned to a valuable sample is 0.02, whereas for a less valuable sample, it is 0.001. Samples with an importance of 0 signify no contribution to the model's accuracy, while values below 0 suggest a detrimental impact, possibly due to incorrect labels or samples lying outside the distribution.

The Shapley value takes into account the contributions of all possible subsets of the training set, offering a more holistic assessment of data importance. However, the accurate computation of the Shapley value based on the defined formula necessitates training $\mathcal{O}(2^N)$ machine learning models, rendering it impractical for complex datasets. As a result, existing methods employ approximate algorithms to estimate the Shapley value. For instance, Ghorbani and Zou [16] introduced two Monte Carlo-based approaches for Shapley value approximation. To expedite evaluation time and enable analysis of large datasets, Jia et al. [24] utilized the K-nearest neighbors (KNN) algorithm to approximate the target learning algorithm, reducing the time complexity to $\mathcal{O}(N \log N)$.

III. Evaluation Setup

In this work, we deploy KNN-Shapley [24] to assess the importance of samples in the training set, which takes a dataset as input and assigns an importance value to each sample in the dataset. This decision is justified by two main considerations.

Firstly, from the perspective of utility, traditional data attribution methods struggle to account for the complex interactions within data subsets. Previous research, as discussed by Gupta and Zou [16], highlights this limitation. Consequently, we adopt Shapley value-based approaches for a more accurate assessment of data importance. We further examine the efficacy of two non-Shapley-based measurement techniques: Leave-one-out (LOO) and the advanced data attribution method, Trak[1] [46]. As evidenced in Figure 20, when comparing the performance of models trained with 5000 samples of varying significance, the accuracy discrepancy is less than 7% for these methods. In contrast, the KNN-Shapley method identifies samples that yield an accuracy difference exceeding 20%, thereby demonstrating its superior capability in accurately quantifying importance, we defer the details to the extended version [65].

Secondly, regarding scalability, most measurement methods are highly computationally inefficient. For instance, employing Leave-one-out to calculate importance values for CIFAR10 necessitates over 80 hours on 8×A100 GPUs. Shapley value-based methods are generally *more demanding*. In Jia et al.'s work [25], the authors provide a runtime demonstration showing that existing measurement methods, except for KNN-Shapley, do not scale efficiently to large datasets, even as CIFAR10.

Furthermore, the comprehensive evaluations conducted by prior studies [25] consistently underscore the effectiveness and accuracy of KNN-Shapley. Therefore, considering both utility and scalability, KNN-Shapley emerges as the sole feasible method for conducting experiments.

Datasets: Our evaluation encompasses three widely-used benchmark datasets, namely CIFAR10 [1], CelebA [36], and TinyImageNet [2]. CIFAR10 comprises a collection of 60,000 colored images evenly distributed across ten classes, representing common objects encountered in everyday life, including airplanes, birds, and dogs. CelebA is a large-scale face dataset that encompasses over 40 annotated binary attributes. To ensure balance in our analysis, we follow previous works [43], [49], [76], [35] that select the three most balanced attributes (Heavy Makeup, Mouth Slightly Open, and Smiling) to create an 8-class (2^3) classification task. Note that our findings are not dependent on this specific attribute selection, as validated in Appendix A. Additionally, our evaluation incorporates TinyImageNet, which constitutes a subset of the ImageNet dataset. It encompasses 200 distinct object classes, each with 500 training images. We further validate the generalizability of our conclusions across different modalities, with detailed information deferred to Section IX.

A. Learning Characteristic

In order to gain a deep understanding of the disparities between high and low importance data, we delve into the learning characteristics, such as loss, associated with these samples. To the best of our knowledge, our study represents the first endeavor to investigate the learning characteristics of samples with varying degrees of importance, diverging from the conventional focus solely on their contribution to the final performance.

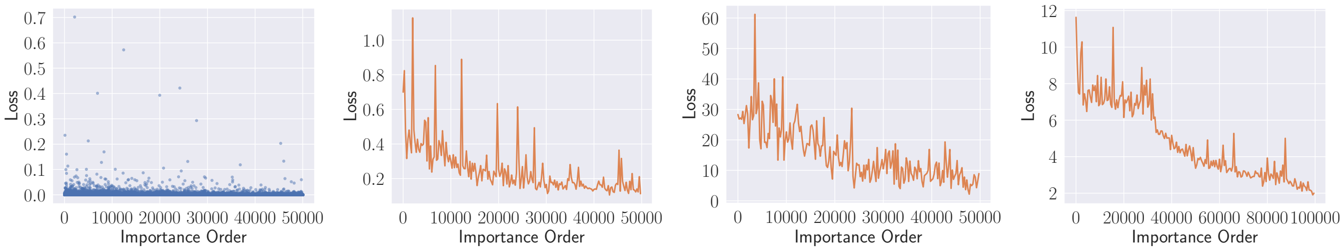
To quantify these learning characteristics, we initially train a model using the complete dataset, comprising both high and low importance data. Subsequently, we compute the loss for each individual data point and explore the correlation between the loss and its corresponding importance value.

In Figure 1a, we present a visual representation of the relationship between loss and importance value. The x-axis represents the importance order of a sample in the dataset, with 1 denoting the lowest importance and 50000 representing the most valuable data. Initially, it may seem that there is no discernible pattern between loss and importance value, as both low and high importance samples can exhibit either low or high loss. However, upon further analysis, we statistically observe that higher importance samples tend to demonstrate lower loss, as depicted in Figure 1b, Figure 1c, and Figure 1d.

To arrive at this conclusion, we divide the samples into 200 bins based on their importance value. For instance, the lowest 1 to 250 samples are categorized into bin one, 251 to 500 are allocated to bin two, and so forth. For each bin, we calculate the sum of the losses and plot these 200 data points to generate the final curve. Despite some fluctuations observed in the curve, it is evident that valuable samples tend to exhibit lower loss. This finding aligns with our expectations, as lower loss signifies greater representativeness, thereby facilitating easier learning and enhancing their contribution to the overall utility of the model.

Having established the effectiveness of importance assignment and gained preliminary insights into the learning

[1] Trak quantifies the influence of each sample on specific test samples within a dataset. To adhere to the established definition of data importance, we calculate the average influence exerted by each sample across the entire test dataset as the importance of each sample.



(a) Loss Distribution (b) CIFAR10 (c) CelebA (d) TinyImageNet

Fig. 1: Relationship between loss and importance value. Low importance samples statistically have higher losses.

characteristics, we proceed to conduct representative machine learning attacks to investigate the impact of data importance in such attacks. Our experimental investigations are carried out using the ResNet18 architecture, and in Section IX, we demonstrate the generalizability of our conclusions to different architectures.

IV. Membership Inference Attack

Membership Inference Attack (MIA) [34], [55], [50], [12], [31], [70] is a prominent privacy attack utilized to determine whether a specific data sample belongs to a training dataset. This attack is widely employed to assess the privacy of training data due to its simplicity and broad applicability.

In the attack scenario, the adversary $\mathcal{A}$ is granted access to a target model and is tasked with determining the membership status of a given data sample (x, y). Formally, the membership inference attack can be defined as a security game, referred to as Membership Inference Security Game, which is described as follows:

Definition IV.1 (Membership Inference Security Game [7])**.** The game proceeds between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$:

1) The challenger samples a training dataset $D \leftarrow \mathbb{D}$ and trains a model $f_\theta \leftarrow \mathcal{T}(D)$ on the dataset D.
2) The challenger flips a bit b, and if $b = 0$, samples a fresh challenge point from the distribution $(x, y) \leftarrow \mathbb{D}$ (such that $(x, y) \notin D$). Otherwise, the challenger selects a point from the training set $(x, y) \overset{\$}{\leftarrow} D$.
3) The challenger sends (x, y) to the adversary.
4) The adversary gets query access to the distribution $\mathbb{D}$, and to the model f_θ, and outputs a bit $\hat{b} \leftarrow \mathcal{A}^{\mathbb{D},f}(x, y)$.
5) Output 1 if $\hat{b} = b$, and 0 otherwise.

The adversary $\mathcal{A}$ is provided with auxiliary information about the data distribution $\mathbb{D}$. This allows the adversary to sample a shadow dataset from the same or a similar distribution, which is a common assumption in the existing literature.

The attack accuracy for the adversary is defined as follows:

$$Acc = \Pr_{x,y,f,b}[\mathcal{A}^{\mathbb{D},f}(x, y) = b].$$

To assess the privacy leakage caused by membership inference attacks (MIAs), we employ two metrics commonly used in prior research, focusing on both worst-case and average-case performance:

1) **(Log-scale) ROC Analysis** [7], which focuses on the true-positive rate at low false-positive rates, effectively capturing the worst-case privacy vulnerabilities of machine learning models.
2) **Membership Advantage** [71], [59], defined as

$$Adv = 2 \times (Acc - 0.5).$$

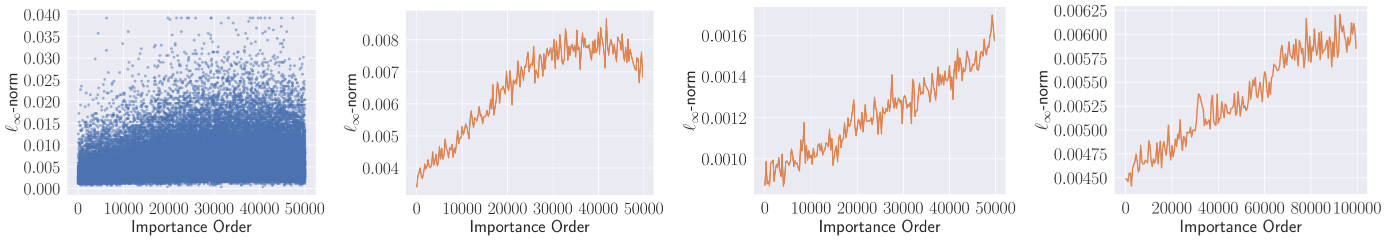
This metric represents the advantage over random guessing, multiplied by 2, providing an average-case measure to gain an overview of the attack's efficacy.

In this work, we investigate four specific membership inference attacks. For the CIFAR10 and CelebA tasks, a training set of 50,000 samples is employed, while for the TinyImageNet task, we utilize a training set of 100,000 samples to construct the target model.

To assess the membership status of samples, we first adopt a methodology based on previous research [12], [31] that considers the distance to the decision boundary as a reflection of membership status. Specifically, they claim that samples located near the decision boundary are more likely to be non-members, whereas samples positioned in the central region of the decision area are more likely to be members.

We calculate the distance to the decision boundary for all samples in the training dataset. Specifically, for each sample, we iteratively perturb it using Projected Gradient Descent (PGD) with a small step size until it is classified into a different class. Subsequently, we compute the distance between the perturbed sample and its original counterpart. In this analysis, the distance is measured using the ℓ_∞ norm, and we find consistent results across different norms such as ℓ_1 and ℓ_2, as evidenced by the corresponding findings presented in the extended version [65].

Our initial visualization focuses on examining the distances of samples with different importance values. Similar to the observation made regarding the distribution of loss values in Section III-A, no direct relationship is discernible between the importance value and the distance to the decision boundary. Notably, samples with similar importance values may exhibit substantial differences in their distances to the decision boundary. In contrast, we further analyze the statistical characteristics of these samples, as performed in Section III-A and present the "group distance" in Figure 2b, Figure 2c, and Figure 2d.



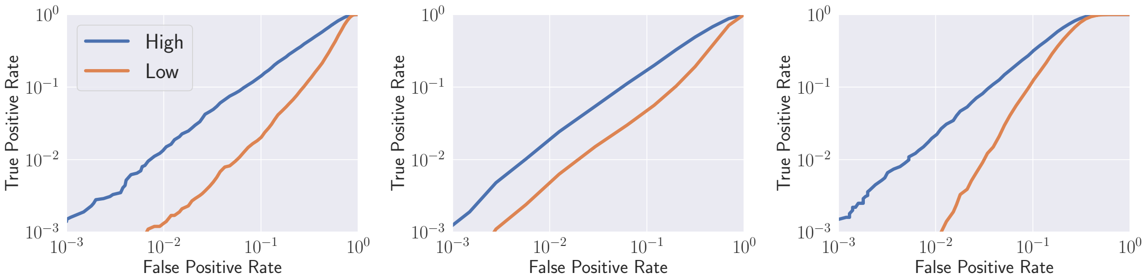
(a) Distance Distribution

(b) CIFAR10

(c) CelebA

(d) TinyImageNet

Fig. 2: Relationship between distance to the decision boundary and importance value. Low importance samples are statistically closer to the decision boundary. The distance measured with different norms can be found in the extended version [65].



(a) CIFAR10

(b) CelebA

(c) TinyImageNet

Fig. 3: Log-scale ROC curve: membership inference attack based on the distance to the decision boundary. High importance samples exhibited substantially higher true-positive rates, particularly in the low false-positive rate region. Results with different norms can be found in the extended version [65].

The results reveal that low importance samples are statistically closer to the decision boundary, which aligns with the previous conclusion that low importance samples tend to have higher loss compared to high importance samples.

We follow the same procedure to derive the distance for samples in the testing dataset and launch the membership inference attack based on the distance. We identify 10,000 samples with the highest importance value as the high group, and an equivalent number of samples with the lowest value as the low group. The resulting ROC curves, depicted in Figure 3, are presented on a logarithmic scale to compare the performance between these two groups.

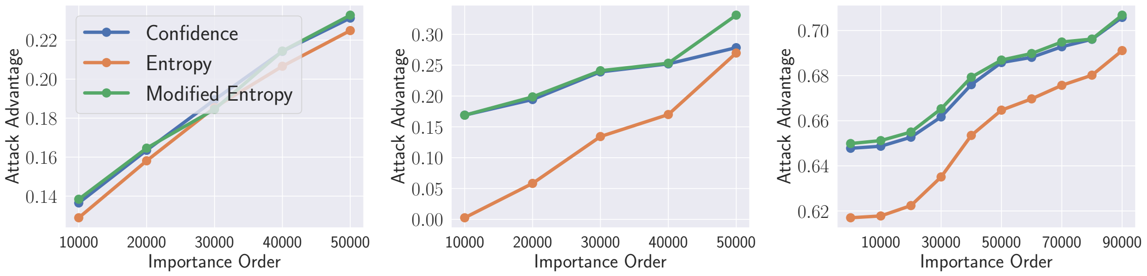
From the figures, we observe significant differences in the behavior of high importance samples and low importance samples, particularly in the low false-positive rate area. Specifically, for the CIFAR10 dataset, high importance samples demonstrate a true-positive rate (TPR) $10.2\times$ higher than low importance samples at a low false-positive rate of 1%. For the TinyImageNet dataset, the difference is even more pronounced, with high importance samples exhibiting a TPR that is $27.9\times$ higher than that of low importance samples at the same false-positive rate.

These observations provide compelling and empirical evidence supporting the notion that high importance samples are considerably more vulnerable to membership inference attacks, which satisfies our expectation as the importance of data samples can be regarded as the proxy of memorization [13]. These findings thus pose a significant and tangible threat to the safeguarding of high importance data privacy. On the other hand, these findings may also prompt researchers to consider adopting strategic sampling methods for more effective privacy auditing [41], [22].

We further validate the generalizability of this finding across various attack methodologies by conducting experiments with three additional metric-based attacks: *prediction confidence-based* attack [71], [59], *entropy-based* attack [55], [50], and *modified prediction entropy-based* attack [58]. The first two attacks were enhanced by introducing class-dependent thresholds, as demonstrated by Song and Mittal [58].

By grouping the samples based on their importance values in intervals of 10,000 samples (equivalent to the size of the testing dataset), we conducted the aforementioned attacks on these subsets. The membership advantage achieved for each subset is illustrated in Figure 4. Notably, a clear monotonic increase in attack advantage is observed as the importance value increases, establishing a positive correlation between the importance value and the susceptibility of membership inference.

This empirical trend aligns with our expectations. As evidenced by the metrics in Figure 1 and Figure 2, samples with lower importance inherently present greater learning challenges compared to their higher-importance counterparts. Even post-learning, these samples exhibit worse membership metrics compared to those of higher importance. This circumstance



(a) CIFAR10 (b) CelebA (c) TinyImageNet

Fig. 4: Membership advantage: membership inference attack based on three metrics. Attack advantage steadily escalates as the importance value of the samples increases.

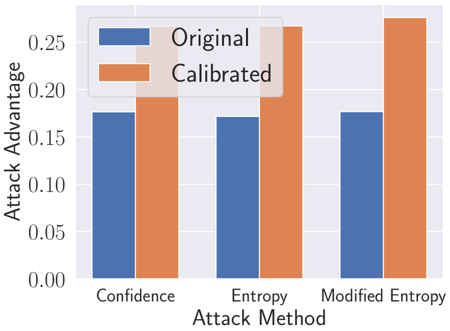


Fig. 5: Incorporation of importance values in calibrating membership inference metrics improves the attack performance, demonstrating the strength of employing sample-specific membership criteria.

renders them challenging to distinguish from non-member samples, especially when certain non-member samples manifest a lower learning difficulty and consequently exhibit better metrics compared to the more challenging member samples.

Drawing from this insight, one potential strategy to enhance the efficiency of membership inference attacks is to compare each sample with others of comparable difficulty. A pragmatic approach to actualize this entails introducing sample-specific criteria. Rather than employing a uniform threshold across the entire testing dataset, such criteria should intricately correlate with the sample's characteristics, with their designated importance level serving as a robust quantitative index to reflect this alignment.

In this study, we initiate an exploration into the feasibility of such an approach. To seamlessly integrate our method into the existing metric-based framework, we introduce a sample-specific threshold in a consistent manner: while maintaining a uniform threshold for the dataset, we modify the membership metrics by incorporating an importance-related term:

$$\texttt{CaliMem}(x) = \texttt{OriMem}(x) + k \times \texttt{Shapley}(x)$$

Here, $\texttt{OriMem}(x)$ denotes the conventional membership metric, including elements such as confidence, entropy, and modified entropy. The term $\texttt{Shapley}(x)$ signifies the importance value attributed to the specific sample x, and $\texttt{CaliMem}(x)$ represents the recalibrated membership metric.

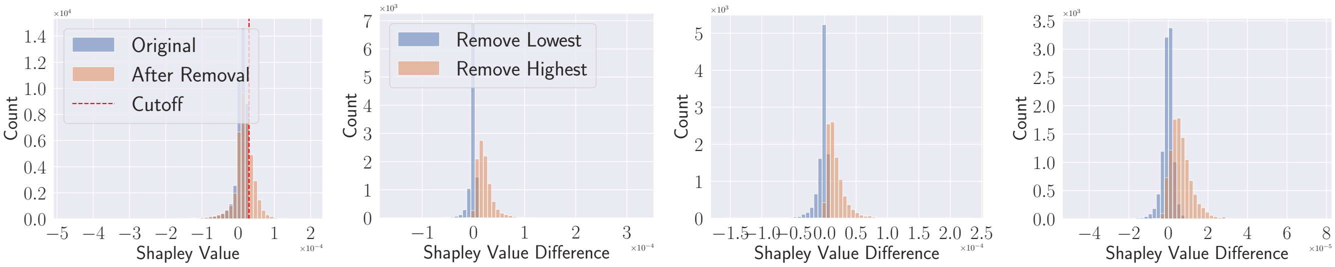
As a proof of concept, we empirically determine the hyperparameter k in an exploratory manner, adjusting its magnitude until optimal performance is attained. The experimental outcomes, depicted in Figure 5, illustrate that the incorporation of importance calibration notably enhances the efficacy of metric-based attacks. Nevertheless, it is pertinent to acknowledge that identifying the optimal hyperparameter and devising more refined methods for integrating importance values warrant further investigation.

We also emphasize that this improvement does not necessitate additional requirements compared to standard attacks; specifically, the adversary does not need full access to the training dataset to obtain importance values. Although importance values cannot be calculated for single samples, most membership inference attacks assume access to a shadow dataset. We validate the feasibility of our approach using a shadow dataset. Specifically, we randomly select 1,000 samples from the CIFAR10 dataset to calculate their importance value, we first calculate the importance value for all samples using the whole CIFAR10 dataset as the ground truth. Then we assume the adversary can only access a shadow dataset containing 10,000 samples, and calculate the importance value for each sample using only the shadow dataset. We found a correlation coefficient of 0.957 between these values, indicating that using the shadow dataset could provide a good approximation.

A. Privacy Onion Effect

Carlini et al. [9] have identified the onion effect of memorization, which refers to the phenomenon wherein "*removing the layer of outlier points that are most vulnerable to a privacy attack exposes a new layer of previously-safe points to the same attack.*" Their research demonstrates this effect by removing samples that are at the highest risk of being compromised through membership inference, resulting in formerly safe samples becoming vulnerable to the attack.

Building upon the insights from the preceding section, our empirical findings confirm a positive correlation between membership inference vulnerability and data importance. This prompts an intriguing question: Does this effect reflect in the importance values assigned to the data? Put differently, when



(a) Distribution Shift

(b) CIFAR10

(c) Celeba

(d) TinyImageNet

Fig. 6: The privacy onion effect can be extended to the importance distribution. For Figure 6a, we remove all examples with importance values larger than the red line. Samples that remain after removal have their importance value increase past the red line. Subsequent figures confirm that removing highly important samples elevates the significance of the remaining set, while the removal of less significant samples has no such effect, ruling out dataset size influence.

high importance samples are removed from a dataset, do previously designated low importance samples gain significance?

To avoid ambiguity, it is imperative to understand that the term "importance" in this context is not subjective or relative. The removal of high importance data points does not inherently increase the importance of those initially deemed low importance. Furthermore, it is conceivable for a dataset to exclusively consist of low importance samples. This clarification is indispensable; otherwise, the studied question may seem trivial.

Our results indicate that removing important samples indeed makes samples previously considered unimportant gain importance. Specifically, upon removing 10,000 data points with the highest importance scores, we recalculated the importance for the remaining samples. As depicted in Figure 6a, this removal led to a noticeable redistribution in data point importance, with previously low important data points now being assigned greater significance.

However, a note of caution is warranted in interpreting this result. The removal of a substantial number of samples (10,000 in this context) might introduce a baseline drift. Therefore, attributing the observed importance augmentation solely to the exclusion of important samples might be premature. To further validate our findings, we executed controlled experiments wherein we systematically excluded either the most or least significant data points. This approach mitigates potential biases stemming from dataset size discrepancies. To emphasize the impact of these exclusions, we quantified the importance value discrepancies for data points ranked between 10,000 and 20,000 in descending order of importance in the original dataset, as they remained for both removal procedures.

Our results, as visualized in Figure 6b, Figure 6c, and Figure 6d, underscore the pronounced disparities in how the exclusion of high importance versus low importance data samples influences the remaining dataset's importance distribution. Using CIFAR10 as an illustrative case, removing the most significant data points caused 99.14% of the remaining data points to be reevaluated as more important. In contrast, removing the least significant data points led to a 45.88% decrease in importance for the affected data points. These findings robustly support our hypothesis that data points previously deemed of lesser importance assume greater significance when high importance data points are excluded, and such a conclusion cannot be attributed to dataset size variation.

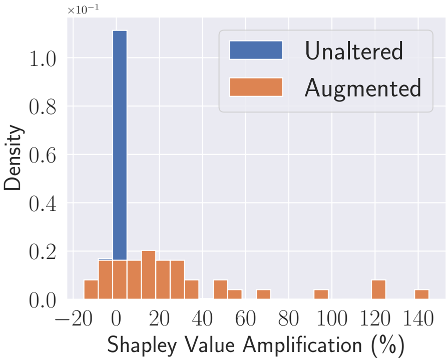
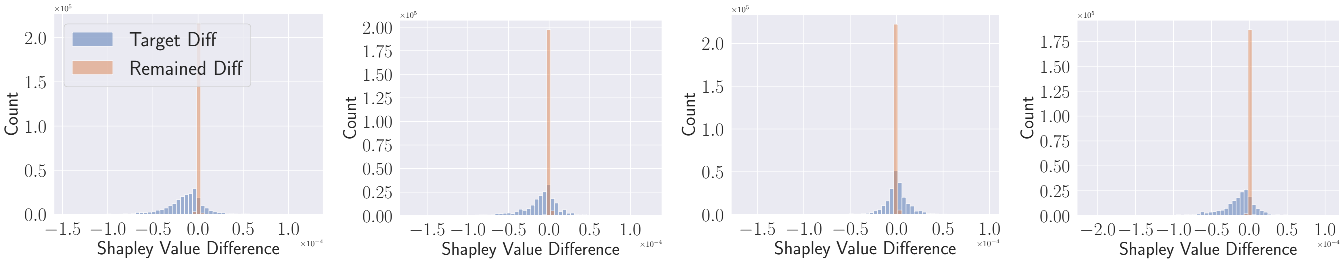


Fig. 7: Impact of sample duplication on importance values. Following the duplication process (16 duplications per sample), 45 out of 50 target samples showed a marked increase in importance, averaging a 53.02% rise, while the importance of control samples (unduplicated) remained relatively stable.

B. *Actively Modify Sample Importance*

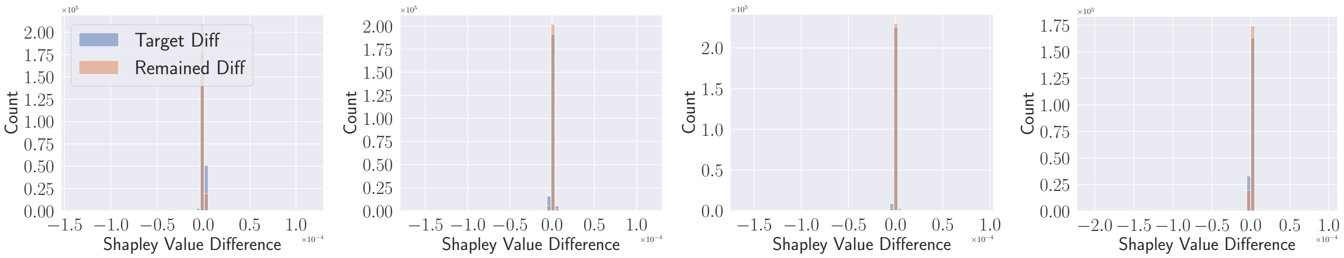
As discussed in the previous section, altering the dataset can influence the importance of samples. Given the observed linkage between membership vulnerability and importance value, an interesting question arises: can we actively use our findings to design more advanced attacks by modifying the importance of target samples?

However, directly altering sample importance is challenging due to the absence of a standardized method or framework. In this section, we explore an ad-hoc approach aimed at increasing the importance of target samples. Specifically, we select a set of target samples and duplicate each one multiple times with consistently incorrect labels. This strategy intuitively heightens the influence of these samples by causing the model to consider them as "outliers" due to the prevalence of incorrect duplicates. We tested this idea by duplicating 50 target samples 16 times and reassessing their importance



(a) ColorJitter (b) GrayScale (c) HorizontalFlip (d) VerticalFlip

Fig. 8: Comparison of importance values before and after replacing 1,000 samples with their augmented versions using ColorJitter, Grayscale, HorizontalFlip, and VerticalFlip techniques. The plot illustrates that augmentation caused variable changes in importance, with some samples gaining and others losing importance.



(a) ColorJitter (b) GrayScale (c) HorizontalFlip (d) VerticalFlip

Fig. 9: Analysis of the impact on importance values when 1,000 augmented samples are added to the original dataset, compared to a baseline scenario where the same 1,000 samples were duplicated. The figure demonstrates that the presence of both original and augmented samples had minimal effect on the importance of the original samples, showing no significant differences from simple duplication.

values. As shown in Figure 7, the duplicated samples generally exhibited an increase in importance. Specifically, 45 of the 50 target samples experienced an increase in importance, with an average increase of 53.02%, while the importance of the unaltered samples remained nearly constant.

This approach is exactly the membership poisoning attack proposed by Tramèr et al. [61], where they comprehensively demonstrated the method's efficiency. This indicates that increasing the importance of samples can be a practical approach to enhance their vulnerability. By revisiting their technique from a data importance perspective, we highlight that actively modifying sample importance could be a promising strategy for developing sophisticated attack techniques or formulating robust defenses.

C. Data Augmentation

In previous discussions, it has been demonstrated that manipulating data samples can alter their importance values and, consequently, their susceptibility to attacks. Given that data augmentation is the most widely used method for data manipulation, it would be interesting to ask: does data augmentation affect the importance of a sample?

In our study, we examined the impact of four data augmentation techniques—ColorJitter, Grayscale, HorizontalFlip, and VerticalFlip—under two specific scenarios. In both scenarios, we selected 1,000 samples for augmentation while leaving the rest of the dataset unaltered:

Augmented Versions Only: In this scenario, we aimed to investigate how data augmentation impacts the importance of the augmented samples and the unaltered samples in the dataset. Specifically, we *replaced* 1,000 samples with their augmented versions, recalculated their importance values, and compared these values to the original. As illustrated in Figure 8, we found that the augmented versions had variable effects on importance: some samples gained higher importance, while others lost it. Overall, a slight majority of the samples experienced a decrease in importance following augmentation. However, the augmented samples had a negligible impact on the remaining non-augmented samples.

Original and Augmented Versions: In this scenario, we examined the effect of having both augmented and original versions of certain samples in the dataset. We *added* 1,000 augmented samples to the original dataset and recalculated the importance values for the original samples. To control for dataset size, we also considered a baseline case where the same 1,000 samples were duplicated. As shown in Figure 9, the presence of both original and augmented samples had minimal impact on the importance of the original samples, with no significant differences compared to simple duplication.

We acknowledge that more complex augmentation techniques, such as those using generative models, may have different effects. The exploration of these complex augmentation techniques remains an avenue for future research.

Takeaways: Our findings highlight the vulnerability of high importance samples to membership inference attacks. Significant differences were observed in the behavior of high importance and low importance samples, particularly in the low false-positive rate region, where high importance samples exhibited substantially higher true-positive rates. This emphasizes the necessity of addressing the privacy risks associated with high importance samples and implementing effective safeguards. Simultaneously, it encourages researchers to explore strategic sampling methods to enhance the effectiveness of privacy audits.

The observation also suggests a potential enhancement to membership inference attacks through the introduction of sample-specific criteria. We empirically validate the practicality of using importance values to calibrate membership metrics, thereby enhancing attack efficiency.

Moreover, our findings reveal the "privacy onion effect" within the sample importance distribution, where previously overlooked samples gain importance when key samples are removed. Furthermore, by revisiting an advanced membership poisoning attack from the perspective of data importance, we suggest that actively manipulating sample importance can be a potent strategy for developing sophisticated cybersecurity measures, both offensive and defensive, but finding general manipulating methods needs further investigation.

V. Model Stealing

Model stealing attack [63], [62], [52], [21], [8] differs from membership inference attack as it aims to compromise the confidentiality of the model itself rather than exploiting privacy information about training samples. This type of attack does not have information about the target model's architecture or parameters but seeks to create a surrogate model that emulates the functionality of the target model. Such attacks can be employed by adversaries for various purposes, including monetary gains or as a preliminary step for subsequent attacks [45].

The workflow of a model stealing attack is visualized in Figure 10. The adversary samples data from a specific distribution $\mathbb{D}$ and simultaneously queries the target and surrogate models. To ensure similarity between the surrogate and target models, the adversary optimizes the surrogate model to produce similar outputs $\mathcal{S}(x)$ as the target outputs $\mathcal{T}(x)$. While the attack approach is straightforward, selecting an appropriate query data distribution poses a challenge, as it directly impacts the stolen accuracy and query efficiency. Recent research has explored efficient and data-free methods for launching these attacks [44], [26], [52], yet the question of selecting high-quality samples when the target task is known remains intriguing.

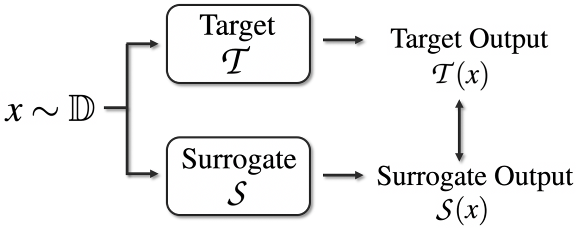


Fig. 10: The workflow of model stealing attack, the adversary leverages the target model to guide the surrogate model.

In this work, we focus on the primary scenario where the adversary can query the target model to obtain corresponding posteriors, while having knowledge of the target task. Specifically, the adversary could query the model with a dataset from the same or a similar distribution. This scenario has practical applications, such as creating a surrogate model to facilitate further attacks or to save on labeling costs. We limit our discussion to this primary scenario and do not delve into more advanced model stealing techniques that focus on reducing the dataset assumption, as our interest lies in understanding how different data interact with the model stealing process.

Our goal is to investigate whether query samples with different importance values exhibit varying efficiency in stealing models. We explore two settings in our experiments. First, we launch the attack using query data from the same distribution as the target model trained on. For example, if the target model is trained on CIFAR10, we employ CIFAR10 data to query the model. The second scenario involves using data from different distributions, specifically CelebA and TinyImageNet, to query the CIFAR10 model. We choose accuracy and query budget as the metrics to evaluate the success of the attack, using less query budget to achieve higher accuracy denotes better attack performance.

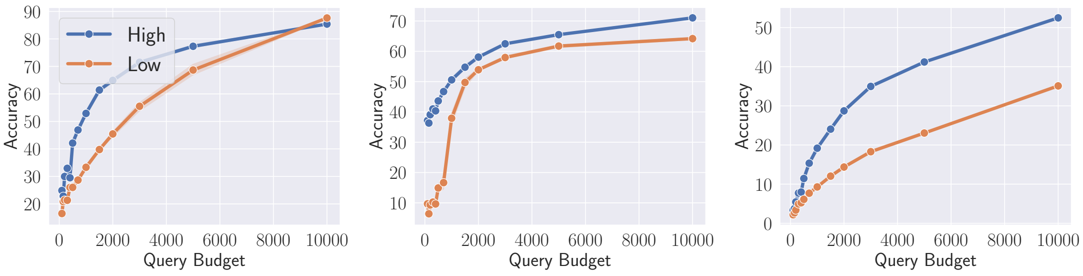
A. *Same Distribution Query*

Three target models were trained using a standard training procedure, resulting in testing accuracies of 95.15% for CIFAR10, 79.05% for CelebA, and 65.01% for TinyImageNet. After training the target models, our attack solely interacts with the target models through their outputs without accessing or reading their parameters.

To initiate the attack, we establish a query budget ranging from 100 to 10,000. Once the query budget is determined, we prioritize collecting high importance data until the budget is exhausted, and the same principle applies to the collection of low importance data.

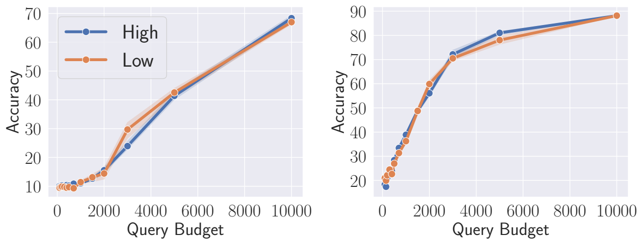
The attack results are illustrated in Figure 11, highlighting the superior efficiency of high importance samples in the model stealing process. For instance, when the query budget is set to 1000, high importance data steal a CIFAR10 model with 53.77% accuracy, which is 1.6× higher than the model stolen by low importance data (33.29%). This trend holds true for the other two datasets as well. Taking TinyImageNet as an example, when the query budget is 1000, high importance data yield a model accuracy of 19.25%, whereas low importance data only result in a model accuracy of 9.25%, exhibiting a notable 2.1-fold disparity.

One plausible explanation for this difference may arise from variations in class balance, given that the query sets are chosen based on sample importance. It is conceivable that the low importance query set may lack samples from certain



(a) CIFAR10 (b) CelebA (c) TinyImageNet

Fig. 11: Model stealing attack that queried with data from the same distribution. High importance samples exhibit greater efficiency in stealing models when the target model is trained on the same distribution as the query distribution.



(a) CelebA (b) TinyImageNet

Fig. 12: Model stealing attack that queried with data from different distributions. The target model is trained on the CIFAR10 task. Results show that importance does not transfer between different tasks.

classes, thereby resulting in suboptimal performance. Prior research has suggested that a more balanced data distribution could potentially improve model stealing performance [52], [33]. However, upon examining the data distribution for both high and low significance samples, we observed no significant disparities. For example, in the case of CIFAR10, the entropy values for the top-10,000 high importance and low importance distributions were 3.282 and 3.245, respectively. Even when considering 1000 samples, the corresponding entropy values were 3.161 (high importance) and 3.229 (low importance). For context, a perfectly uniform distribution has an entropy of 3.322. This indicates that both the high and low importance subsets closely approximate a uniform distribution. Such findings reinforce our assertion that high importance data can indeed augment model stealing performance, mitigating concerns related to distributional biases.

B. Different Distribution Query

The previous section highlighted the enhanced efficiency of high importance samples in stealing models trained on the same task. However, it remains uncertain whether this efficiency persists when the target model is trained using a different dataset or task, and whether importance values can be transferred across tasks.

To investigate, we conducted experiments involving query data that differed from the distribution used to train the target model. Specifically, we employed a CIFAR10 model as the target model and queried it with the CelebA and TinyImageNet datasets.

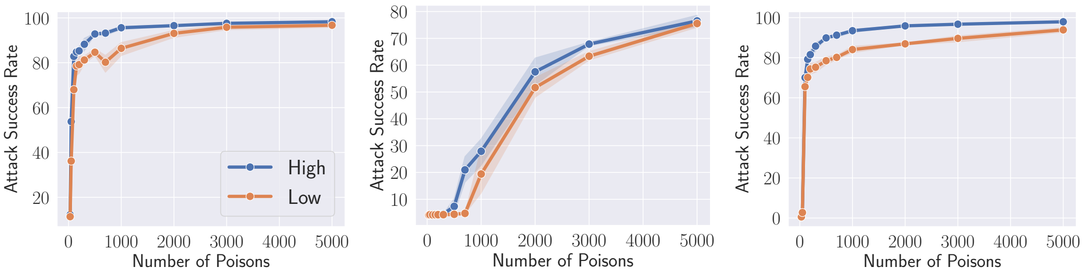
Interestingly, as depicted in Figure 12, we observed that the advantage of high importance samples disappeared in this cross-task scenario. When the query budget was consistent, the stolen accuracy for both high and low importance samples was comparable. This suggests that samples deemed important for one task may not transfer effectively to arbitrary tasks.

Takeaways: Our findings demonstrate that high importance samples exhibit greater efficiency in stealing models when the target model is trained on the same distribution as the query distribution. Importantly, this enhanced efficiency cannot be solely attributed to distribution bias. This suggests that adversaries, when aware of the target task, can employ high importance samples to optimize attack performance with a reduced query budget. However, this conclusion does not hold when the target task differs from the query distribution. Consequently, this implies that selecting a group of high importance samples as a “universal” query set for efficient model stealing attacks, regardless of the target task, is not feasible.

VI. Backdoor Attack

Backdoor attack [17], [32], [49] is a training-time attack that involves actively interfering with the training process to manipulate the resulting model. Its primary objective is to introduce malicious behavior into the model, making it behave like a benign model for normal inputs. However, when a specific trigger is detected, the backdoored model intentionally misclassifies the input to a predetermined class. This type of attack can have severe consequences, such as compromising the integrity and reliability of the model, leading to potential security breaches, data manipulation, or unauthorized access to sensitive information.

Despite the severe consequences that a backdoor attack may cause, the attack itself is relatively easy to achieve by poisoning the training dataset, thereby posing an even stronger threat. For instance, a straightforward attack approach called BadNets [17] adds a fixed trigger to a portion of the training dataset, resulting in a perfect attack where almost all triggered samples are misclassified into the target class, while the accuracy on the original task remains largely unaffected.



(a) CIFAR10

(b) CelebA

(c) TinyImageNet

Fig. 13: Relationship between attack success rate and the poisoning rate, high importance samples enhance the efficiency of the poisoning process, particularly when the poisoning rate is small.

In the context of backdoor attacks, the poison rate plays a critical role as it directly influences the effectiveness and concealment of the attack. A higher poison rate can lead to an increased attack success rate, but it also raises the risk of detection since a large number of samples need to be modified. Conversely, a lower poison rate may offer better concealment, but it may not achieve optimal attack performance. Additionally, there are situations where the adversary can only control a small set of samples, making it impossible to poison a large number of samples to achieve the attack. Consequently, the problem of backdooring a model with a limited poison rate becomes an interesting and challenging research question.

In this section, we conduct empirical investigations to explore whether poisoning samples with different importance levels influences the attack performance under the same poison rate. We utilize two metrics to evaluate the attack performance:

1) **Accuracy**. This metric assesses the deviation of the backdoored model from the clean model. We measure the performance of the backdoored model on the clean dataset, and a successful attack should result in accuracy close to that of the clean model, making it difficult to detect.
2) **Attack Success Rate (ASR)**. This metric evaluates the functionality of the backdoored model and is measured on the triggered dataset. A desirable backdoored model should exhibit a high ASR, indicating its ability to misclassify all triggered samples into the target label.

By analyzing these metrics, we aim to gain insights into the influence of data importance on the attack performance and further understand the trade-offs between attack effectiveness and concealment in the context of backdoor attacks.

In this part, we adopt the same approach as BadNets to backdoor the model, with hyperparameter details provided in Appendix C. Additionally, we validate the generalizability of our conclusion across five other backdoor attacks—Blend [11], SSBA [30], LF [73], SIG [6], and CTRL [29]—which utilize various trigger patterns or target different learning paradigms, as discussed in the extended version [65].

We present the visualized attack success rate in Figure 13. As depicted in the figure, there is a noticeable increase in the attack success rate as the number of poisons increases. Concurrently, we observe significant differences between poisoning high importance samples and low importance samples. Specifically, poisoning an equal number of high importance samples proves to be more effective in increasing the attack success rate compared to poisoning low importance samples. This phenomenon becomes more pronounced when the poisoning rate is small. For instance, in the case of CIFAR10, with a poisoning size of 50, poisoning high importance data results in a model with an ASR of 54.42%, whereas poisoning low importance data only achieves 37.74%, indicating a $1.44\times$ advantage. Similar trends can be observed across the other two datasets.

However, we also find that when the poisoning rate is large, the difference is not significant. We believe this is due to the trade-off between the importance advantage and the attack upper bound. As the number of poisons increases, the advantage of using high importance data becomes more evident. However, achieving the optimal ASR for the backdoor attack does not require a large amount of data. Generally, poisoning approximately 10% of the dataset is sufficient. Therefore, as the number of poisons increases, the gap between high importance and low importance samples is reduced. Nevertheless, it is still observed that poisoning high importance samples requires poisoning fewer samples to achieve its optimal ASR.

In scenarios where adversaries have limited access to data, determining the true importance of samples can be challenging, which impacts the feasibility of selectively poisoning high importance samples. In this case, we empirically demonstrate that calculating the importance value using just a fraction of the training set can provide a good approximation of the true importance. For example, with just 2% of the CIFAR10 data available, the computed importance values correlate strongly with those derived from the entire dataset, achieving a correlation coefficient of 0.811 ± 0.016. The accuracy of these approximations improves with more data: with 5% of the data, the correlation coefficient rises to 0.899 ± 0.006, and with 20% of the data, it exceeds 0.96. These results demonstrate that even with limited data access, it is feasible to closely estimate the importance of samples, facilitating effective attack planning under realistic constraints.

Additionally, our investigation into the impact on clean accuracy reveals no significant trends suggesting that poisoning samples of differing importance levels affects clean accuracy.

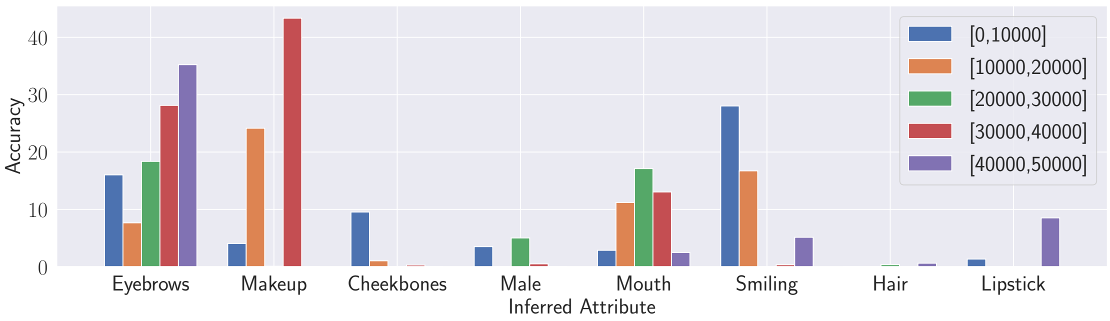


Fig. 14: Attribute inference attack performance on different attributes, no significant correlation between attribute inference attacks and data importance is observed. These confirm our hypothesis that the importance of data samples is context-dependent and can vary based on the specific task at hand.

In both scenarios, the influence on clean accuracy remains below 2%, indicating the concealment of the backdoor attack. Due to space constraints, detailed results are presented in the extended version [65].

Takeaways: Our experimental results demonstrate that poisoning high importance samples enhances the efficiency of the poisoning process, particularly when the poisoning rate is small. This insight offers valuable guidance for developing attack strategies aimed at compromising models with restricted data accessibility. Beyond refining trigger patterns for effective injections, prioritizing the poisoning of high importance samples emerges as a promising approach. On the other hand, the influence on clean accuracy does not yield a definitive conclusion, as poisoning either type of data has a limited impact on clean accuracy.

VII. Attribute Inference Attack

Attribute inference attack is a privacy attack that aims to infer sensitive attributes that are not directly related to the original task of a machine learning model. For instance, a model trained to predict age from profile photos may unintentionally learn to predict race as well [56], [39], [57]. This type of attack has significant implications for privacy and fairness, as the inadvertent leakage of sensitive attributes can have far-reaching consequences, including the violation of privacy rights, potential discrimination, and the undermining of trust in machine learning systems.

In this work, we focus on a commonly considered attack scenario as depicted in Figure 15, where the adversary exploits the embeddings of a target sample obtained from the target model to predict its sensitive attributes [2] To perform attribute inference, the adversary assumes auxiliary information about the training dataset and collects a shadow dataset from similar distributions. They train a shadow model to mimic the behavior of the target model and use the embeddings and sensitive attributes to train an attack classifier.

In this section, we investigate the impact of data importance on the CelebA dataset, which contains several attributes that can be inferred. We categorize the samples into five groups, each comprising 10,000 samples, based on their importance values ranging from low to high. Following this categorization, we train five models using these groups as target models.

To perform the attack, we utilize 10,000 samples, disjoint from the 50,000 training samples, to train a shadow model. This shadow model is employed to generate datasets for training the attack model, where the inputs are embeddings, and the associated sensitive attributes serve as labels. We train a two-layer fully connected network as the attack model, which is then utilized to infer the sensitive attribute from the embeddings.

To evaluate the attack performance, we utilize relative accuracy as the metric, comparing the accuracy against a random guessing baseline that varies for different attributes due to the uneven distribution of the CelebA dataset.

The experimental results, as shown in Figure 14, reveal no significant connection between data importance and the success of attribute inference attacks. For instance, the “Arched Eyebrows” attribute is easily inferred for high importance samples, while only low importance samples can be inferred for the “High Cheekbones” attribute. Furthermore, the vulnerability to attribute inference for the “Mouth Slightly Open” attribute is most prominent among samples with middle importance values. These results demonstrate that there is no significant correlation between attribute inference attacks and data importance.

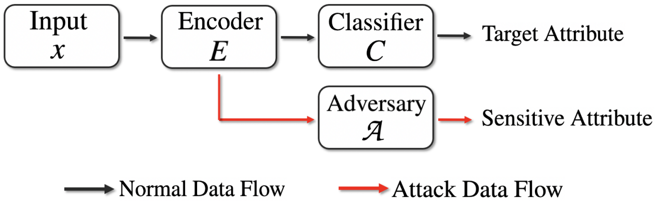
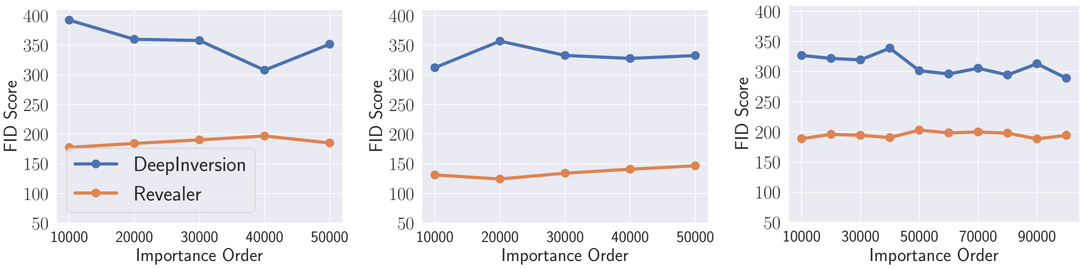


Fig. 15: The attack scenario for attribute inference attack. The adversary can get the embeddings and aims to infer sensitive attributes based on the information encoded in embeddings.

[2] We acknowledge that there exists a separate line of research on attribute inference attacks targeting *tabular data* [38], [23], [15], which primarily aims to *reconstruct* missing attribute values in original records. Given that these attacks employ different technical methodologies and pursue distinct objectives, our conclusions may not necessarily apply to such work.



(a) CIFAR10

(b) CelebA

(c) TinyImageNet

Fig. 16: Relationship between reconstruction quality and data importance, reconstruction performance remains steady regardless of the importance level of the data samples.

One possible explanation for these results is that the importance value of data samples may vary depending on the prediction task. In other words, the significance of certain features or attributes may differ across different prediction tasks. For example, while whiskers may be an important feature for predicting gender, it may hold less importance when predicting income. We further validate our conjecture by visualizing the correlation among importance values assigned to different attributes in the extended version [65]. It indicates that a sample's elevated importance on one attribute may not align with its importance on another attribute.

Takeaways: Our findings indicate that there is not a straightforward correlation between the significance of data samples and the performance of an attack, aligning with our initial hypothesis. A pivotal insight from this section demonstrates that the importance of data samples is context-dependent and can vary based on the specific task at hand, which resonates with our earlier discovery in Section V-B.

VIII. Data Reconstruction Attack

Data reconstruction attack [75], [72], [14], [69], [48] refers to recovering the target dataset with limited access to the target model, with the aid of additional knowledge possessed by the adversary. While data reconstruction attack shares similarities with membership inference attack, there are significant differences that make data reconstruction a stronger attack. Specifically, membership inference operates at the sample level, determining the membership status of individual samples. In contrast, data reconstruction is a dataset-level attack aimed at extracting the entire training dataset. This distinction necessitates different technical approaches for data reconstruction.

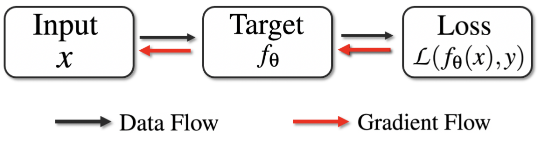


Fig. 17: The workflow of a basic data reconstruction attack, the adversary optimizes the input to maximize the likelihood of the target class.

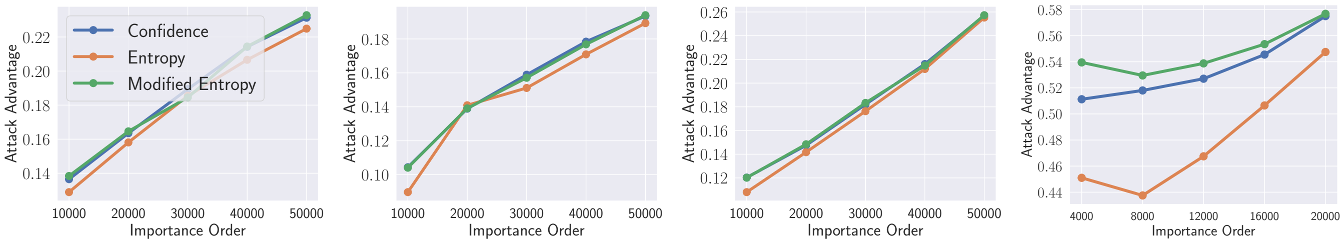
In this work, we employ two data reconstruction attacks, namely DeepInversion [72] and Revealer [75], to investigate the influence of data importance on the reconstruction process. These attacks are based on the optimization of input samples, as illustrated in Figure 17. Specifically, given a target class y, both methods initialize a sample x and iteratively update it to maximize the likelihood or probability of belonging to that class while keeping the model parameters fixed. This optimization process is guided by the following loss function:

$$\min_x \mathcal{L}(f_\theta(x), y)$$

DeepInversion leverages statistical information encoded in the batch normalization layer to enhance the quality of reconstructions, while Revealer employs a Generative Adversarial Network (GAN) to generate high-quality reconstructions.

To investigate the impact of data importance on the performance of data reconstruction attacks, we partition the samples into groups of 10,000 based on their importance values, ranging from low importance to high importance. Subsequently, we train one target model for each sample group, resulting in a total of five models for CIFAR10 and CelebA datasets, and ten models for TinyImageNet. We then apply two reconstruction attacks on each of them. We leverage Fréchet Inception Distance (FID) to measure the similarity between reconstructed samples and the training samples, given its established utility in evaluating the quality of generated distributions [75], [72], [60]. A smaller FID denotes better reconstruction quality. For each target model, we generate 10,000 reconstructions, matching the size of the training dataset. Subsequently, we calculate the FID score, quantifying the discrepancy between the reconstructions and the corresponding training dataset.

The findings presented in Figure 16 suggest that there is no significant distinction between high and low importance data samples in terms of data reconstruction. Taking CIFAR10 as an example, DeepInversion exhibits a maximum deviation of only 13.02% compared to the mean value, indicating a consistent performance. Similar results are observed with Revealer, where the maximum deviation is merely 5.35% compared to the mean value. Moreover, this consistent performance extends to more complex datasets. For instance, in the case of CelebA dataset, the maximum deviation is less than 8.27%, while for TinyImageNet, the deviation is less than 4.01%. These findings suggest that the reconstruction performance remains steady regardless of the importance level of the data samples.



(a) ResNet18 (CIFAR10)

(b) MobileNetV2 (CIFAR10)

(c) ResNet50 (CIFAR10)

(d) MLP (Purchase)

Fig. 18: Relationship between membership inference attack advantage and data importance. Results show that our conclusion can be generalized to different model architectures and data modalities.

IX. Transferability Study

In order to fortify the generalizability of our conclusions, this section investigates the transferability of our findings across various model architectures and data modalities. For the vision modality, we conducted experiments employing two distinct model architectures, namely MobileNetV2 [51] and ResNet50 [19].

To evaluate the transferability to diverse data modalities, we introduced the tabular dataset Purchase-100 [3], consisting of 600 binary features for classifying 100 classes. We focus on the tabular modality considering that existing Shapley methods predominantly support vision and tabular modalities. We utilize a Multilayer Perceptron (MLP) to process the Purchase task, aligning with established practices in prior research [55], [50].

Our experimental results consistently support our conclusions, irrespective of the model architecture and data modality. For example, in Figure 18, the three left figures depict a consistent relationship between the advantage of membership inference attacks and the importance of data across all three model architectures. This trend is also evident when performing the attack based on the distance to the boundary (see results in the extended version [65]). Additionally, Figure 18d illustrates that this conclusion holds for the tabular modality. Although slight fluctuations are observed in the low importance area, the overall picture demonstrates a consistent relationship between importance and membership vulnerability, aligning with the conclusions drawn from the vision modality.

Furthermore, this conclusion extends to other attack types, such as model stealing and backdoor attacks. Due to space constraints, we defer full results to the extended version [65]. These findings reaffirm the consistent impact of data importance across different attack scenarios, underscoring the generalizability of our observations.

X. Limitations and Future Work

While our research provides valuable insights into the relationship between data importance and vulnerability to specific attacks, several limitations exist that warrant further investigation. Our study focuses on a specific set of attacks. Although these are important, they may not cover the entire spectrum of potential threats. Other types of attacks could exhibit different relationships between data importance and vulnerability, and understanding how these various attacks interact with data importance remains an open area for exploration.

Extending our findings to Large Language Models (LLMs) presents substantial challenges despite their promising advancements. The primary obstacle is the computational cost associated with calculating importance values. To manage this burden, current methods often resort to computationally lighter algorithms like KNN for classification tasks. However, it is unclear whether similar computationally efficient approaches can be adapted to approximate auto-regression models, especially since LLMs exhibit unique emergent characteristics when scaled beyond certain thresholds. Additionally, the considerably larger datasets typical of LLMs further complicate the feasibility of extending these methods.

Furthermore, our research does not examine more complex augmentation techniques, such as those utilizing generative models. Future work should investigate whether these advanced techniques affect data importance and vulnerability differently. Additionally, exploring whether there exists a generalizable method to manipulate data importance across various augmentation techniques would be invaluable. To foster further research and collaboration, we have open-sourced our evaluation framework, available at https://github.com/TrustAIRLab/importance-in-mlattacks. This will enable other researchers to examine whether the observed data discrepancies hold for new types of attacks, thereby benefiting the broader community.

XI. Conclusion

In this paper, our research systematically studies the vulnerability of heterogeneous data when confronted with machine learning attacks. Our findings underscore a heightened susceptibility of high importance data samples to privacy attacks, including membership inference attacks and model stealing attacks. Our findings also carry practical implications, inspiring researchers to design more efficient attacks. For example, we empirically showcase the potential enhancement of membership inference attacks through the incorporation of sample-specific criteria based on importance values. Additionally, we demonstrate that our findings can be strategically employed to guide the creation of more advanced attacks through the active manipulation of sample importance.

Acknowledgements

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

References

[1] https://www.cs.toronto.edu/~kriz/cifar.html.

[2] https://www.kaggle.com/c/tiny-imagenet.

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

[4] Anish Agarwal, Munther A. Dahleh, and Tuhin Sarkar. A Marketplace for Data: An Algorithmic Solution. In *Entertainment Computing (EC)*, pages 701–726. ACM, 2019.

[5] Eugene Bagdasaryan and Vitaly Shmatikov. Blind Backdoors in Deep Learning Models. In *USENIX Security Symposium (USENIX Security)*, pages 1505–1521. USENIX, 2021.

[6] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A New Backdoor Attack in CNNS by Training Set Corruption Without Label Poisoning. In *IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019.

[7] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1897–1914. IEEE, 2022.

[8] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. Cryptanalytic Extraction of Neural Network Models. In *Annual International Cryptology Conference (CRYPTO)*, pages 189–218. Springer, 2020.

[9] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramèr. The Privacy Onion Effect: Memorization is Relative. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

[10] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-Enriched Distributional Model Inversion Attacks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 16158–16167. IEEE, 2021.

[11] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *CoRR abs/1712.05526*, 2017.

[12] Christopher A. Choquette Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In *International Conference on Machine Learning (ICML)*, pages 1964–1974. PMLR, 2021.

[13] Vasisht Duddu, Sebastian Szyller, and N. Asokan. SHAPr: An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning. *CoRR abs/2112.02230*, 2021.

[14] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1322–1333. ACM, 2015.

[15] Matt Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *USENIX Security Symposium (USENIX Security)*, pages 17–32. USENIX, 2014.

[16] Amirata Ghorbani and James Y. Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning (ICML)*, pages 2242–2251. PMLR, 2019.

[17] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Grag. Badnets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *CoRR abs/1708.06733*, 2017.

[18] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing Training Data from Trained Neural Networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[20] Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. Node-Level Membership Inference Attacks Against Graph Neural Networks. *CoRR abs/2102.05429*, 2021.

[21] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High Accuracy and High Fidelity Extraction of Neural Networks. In *USENIX Security Symposium (USENIX Security)*, pages 1345–1362. USENIX, 2020.

[22] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing Differentially Private Machine Learning: How Private is Private SGD? In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020.

[23] Bargav Jayaraman and David Evans. Are Attribute Inference Attacks Just Imputation? In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1569–1582. ACM, 2022.

[24] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gürel, Bo Li, Ce Zhang, Costas J. Spanos, and Dawn Song. Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms. *Proceedings of the VLDB Endowment*, 2019.

[25] Ruoxi Jia, Fan Wu, Xuehui Sun, Jiacen Xu, David Dao, Bhavya Kailkhura, Ce Zhang, Bo Li, and Dawn Song. Scalability vs. Utility: Do We Have To Sacrifice One for the Other in Data Importance Quantification? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8239–8247. IEEE, 2021.

[26] Sanjay Kariyappa, Atul Prakash, and Moinuddin K. Qureshi. MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13814–13823. IEEE, 2021.

[27] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning (ICML)*, pages 1885–1894. PMLR, 2017.

[28] Yongchan Kwon and James Zou. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 8780–8802. JMLR, 2022.

[29] Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. An Embarrassingly Simple Backdoor Attack on Self-supervised Learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4344–4355. IEEE, 2023.

[30] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible Backdoor Attack with Sample-Specific Triggers. In *IEEE International Conference on Computer Vision (ICCV)*, pages 16443–16452. IEEE, 2021.

[31] Zheng Li and Yang Zhang. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 880–895. ACM, 2021.

[32] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning Attack on Neural Networks. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2018.

[33] Yiyong Liu, Rui Wen, Michael Backes, and Yang Zhang. Efficient Data-Free Model Stealing with Label Diversity. *CoRR abs/2404.00108*, 2024.

[34] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership Inference Attacks by Exploiting Loss Trajectory. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2085–2098. ACM, 2022.

[35] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*, pages 4525–4542. USENIX, 2022.

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning

Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE, 2015.

[37] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 4765–4774. NIPS, 2017.

[38] Shagufta Mehnaz, Sayanton V. Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino. Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models. In *USENIX Security Symposium (USENIX Security)*, pages 4579–4596. USENIX, 2022.

[39] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting Unintended Feature Leakage in Collaborative Learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 497–512. IEEE, 2019.

[40] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization. In *Workshop on Security and Artificial Intelligence (AISec)*, pages 27–38. ACM, 2017.

[41] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2021.

[42] Tuan Anh Nguyen and Anh Tran. Input-Aware Dynamic Backdoor Attack. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020.

[43] Tuan Anh Nguyen and Anh Tuan Tran. WaNet - Imperceptible Warping-based Backdoor Attack. In *International Conference on Learning Representations (ICLR)*, 2021.

[44] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff Nets: Stealing Functionality of Black-Box Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4954–4963. IEEE, 2019.

[45] Nicolas Papernot, Patrick D. McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks Against Machine Learning. In *ACM Asia Conference on Computer and Communications Security (ASIACCS)*, pages 506–519. ACM, 2017.

[46] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing Model Behavior at Scale. In *International Conference on Machine Learning (ICML)*, pages 27074–27113. PMLR, 2023.

[47] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016.

[48] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. In *USENIX Security Symposium (USENIX Security)*, pages 1291–1308. USENIX, 2020.

[49] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic Backdoor Attacks Against Machine Learning Models. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 703–718. IEEE, 2022.

[50] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

[51] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520. IEEE, 2018.

[52] Sunandini Sanyal, Sravanti Addepalli, and R. Venkatesh Babu. Towards Data-Free Model Stealing in a Hard Label Setting. *CoRR abs/2204.11022*, 2022.

[53] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 6103–6113. NeurIPS, 2018.

[54] Lloyd S Shapley. A Value for n-person Games. *Contributions to the Theory of Games*, 1953.

[55] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (S&P)*, pages 3–18. IEEE, 2017.

[56] Congzheng Song and Ananth Raghunathan. Information Leakage in Embedding Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 377–390. ACM, 2020.

[57] Congzheng Song and Vitaly Shmatikov. Overlearning Reveals Sensitive Attributes. In *International Conference on Learning Representations (ICLR)*, 2020.

[58] Liwei Song and Prateek Mittal. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2021.

[59] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 241–257. ACM, 2019.

[60] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting. Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks. In *International Conference on Machine Learning (ICML)*, pages 20522–20545. PMLR, 2022.

[61] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022.

[62] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium (USENIX Security)*, pages 601–618. USENIX, 2016.

[63] Jean-Baptiste Truong, Pratyush Maini, Robert J. Walls, and Nicolas Papernot. Data-Free Model Extraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4771–4780. IEEE, 2021.

[64] Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard S. Zemel, and Alireza Makhzani. Variational Model Inversion Attacks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9706–9719. NeurIPS, 2021.

[65] Rui Wen, Michael Backes, and Yang Zhang. Understanding Data Importance in Machine Learning Attacks: Does Valuable Data Pose Greater Harm? *CoRR abs/2409.03741*, 2024.

[66] Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership Inference Attacks Against In-Context Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.

[67] Rui Wen, Tianhao Wang, Michael Backes, Yang Zhang, and Ahmed Salem. Last One Standing: A Comparative Analysis of Security and Privacy of Soft Prompt Tuning, LoRA, and In-Context Learning. *CoRR abs/2310.11397*, 2023.

[68] Rui Wen, Zhengyu Zhao, Zhuoran Liu, Michael Backes, Tianhao Wang, and Yang Zhang. Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning? In *International Conference on Learning Representations (ICLR)*, 2023.

[69] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, page 225–240. ACM, 2019.

[70] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced Membership Inference Attacks against Machine Learning Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 3093–3106. ACM, 2022.

[71] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[72] Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to Distill:

Data-Free Knowledge Transfer via DeepInversion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8712–8721. IEEE, 2020.

[73] Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective. In *IEEE International Conference on Computer Vision (ICCV)*, pages 16453–16461. IEEE, 2021.

[74] Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, and Yang Zhang. Generated Distributions Are All You Need for Membership Inference Attacks Against Generative Models. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 4827–4837. IEEE, 2024.

[75] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 250–258. IEEE, 2020.

[76] Nan Zhong, Zhenxing Qian, and Xinpeng Zhang. Imperceptible Backdoor Attack: From Input Space to Feature Representation. In *International Joint Conferences on Artifical Intelligence (IJCAI)*, pages 1736–1742. IJCAI, 2022.

[77] Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable Clean-label Poisoning Attacks on Deep Neural Nets. In *International Conference on Machine Learning (ICML)*, pages 7614–7623. JMLR, 2019.

Appendix

A. CelebA Attribute Selection

The CelebA dataset contains 40 binary attributes, which is not suitable for multi-class classification. Therefore, we follow previous works [43], [49], [76], [35] that select the three most balanced attributes (Heavy Makeup, Mouth Slightly Open, and Smiling) to create an 8-class (2^3) classification task.

To validate that our findings are not dependent on this specific attribute selection, we conducted the same experiments using another randomly selected set of attributes (High Cheekbones, Arched Eyebrows, and Wearing Lipstick). We evaluated the performance on membership inference attacks, model stealing, and backdoor attacks. The results are depicted in Figure 19, confirming that our findings are consistent across these different attribute sets. For example, Figure 19a demonstrate that samples with higher importance are more vulnerable to membership inference attack, it also reflects on the worst-case evaluation as illustrated in Figure 19b. The conclusion holds for backdoor and model stealing attack, specifically, with a 1500 query budget, high importance samples can steal a surrogate model with 48% higher accuracy than that stolen by low importance samples.

B. Measurement Hyperparameter

In implementing the KNN-Shapley method, we set the hyperparameter $k = 6$, following the suggestion in the original paper [24]. Further experimentation with $k = 7$ and $k = 8$ indicated that performance remains largely consistent across these settings. Specifically, the correlation between importance values calculated with $k = 6$ and $k = 7$ is 0.9988, and between $k = 6$ and $k = 8$, it's 0.9972, demonstrating the robustness of our results with respect to this hyperparameter.

C. Hyperparameters for Backdoor Attacks

For the three datasets evaluated in our study—CIFAR10, CelebA, and TinyImagenet—a consistent modification was applied to each image: a black square was positioned at the bottom left corner.

The dimension of this black square, or backdoor trigger, varied according to the image sizes of the respective datasets to maintain proportional consistency. Specifically, for the CIFAR10 dataset, with an image resolution of 32×32, the trigger was sized at 2×2. In the case of CelebA, which features larger images with dimensions of 178×218, the trigger's size was increased to 8×8. Lastly, for images from TinyImagenet, which are 64×64 pixels, a 5×5 square was used as the trigger.

D. Related Work

In addition to the attack approach investigated in our work, several methods exist for privacy and security attacks against machine learning models. In the following section, we provide a brief overview of these approaches.

1) Membership Inference Attack: Membership Inference Attacks (MIA) [55], [50], [31], [66], [20], [74] have emerged as a significant threat to privacy in the context of machine learning models. These attacks aim to reveal the membership status of a target sample, i.e., whether the sample was part of the training dataset or not, thereby directly breaching privacy.

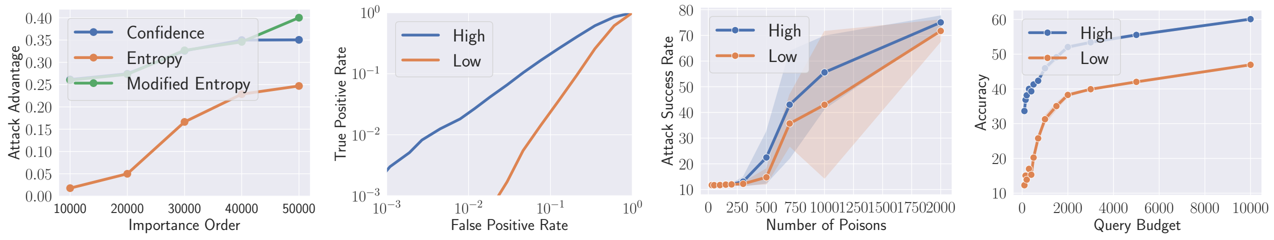
The seminal work by Shokri et al.[55] introduced MIA against machine learning models, wherein multiple shadow models were trained to mimic the behavior of the target model. This attack originally required access to data from the same distribution as the training dataset. However, Salem et al.[50] relaxed this assumption by demonstrating the effectiveness of using only a single shadow model, substantially reducing the computational cost involved.

Subsequent research [12], [31] has explored more challenging settings for MIA. In these scenarios, the adversary only has access to hard-label predictions from the target model. Li and Zhang [31] proposed a method that approximates the distance between the target sample and its decision boundary using adversarial examples, enabling the attacker to make decisions based on this distance.

Recent advancements in MIA have focused on enhancing attack performance. Carlini et al.[7] leveraged the discrepancy between models trained with and without the target sample to improve attack effectiveness. Liu et al.[34] demonstrated the utility of loss trajectory analysis in MIA. Furthermore, Tramèr et al. [61] highlighted the potential of data poisoning, showing that even with access to a small fraction of the training dataset, the attacker can significantly boost the performance of membership inference attacks.

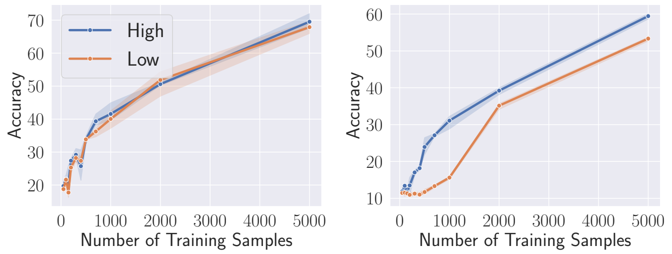
2) Model Stealing Attack: Model stealing attacks [62], [44], [26], [63], [33], [67] aim to extract information from a victim model and construct a local surrogate model. This attack was initially proposed by Tramèr et al.[62], assuming that the adversary has access to a surrogate dataset for stealing the model. Orekondy et al. further advanced this approach by developing a reinforcement learning-based framework that optimizes query time and effectiveness [44].

Recent research has focused on the more stringent data-free setting, where adversaries lack access to any data. In this context, Kariyappa et al. [26] propose MAZE, which employs a generative model to generate synthetic data samples for launching the attack. The generator is trained to maximize disagreement between the victim model and the clone model,



(a) MIA (metric-based) (b) MIA (log-scale) (c) Backdoor ASR (d) Model Stealing

Fig. 19: Attack performance on samples with high and low importance. The results demonstrate that our conclusions are consistent across different sets of selected attributes.



(a) LOO (b) Trak

Fig. 20: LOO and Trak fail to capture the complex interactions between subsets of data, resulting in suboptimal sample importance identification performance.

requiring the gradients from the victim model. To approximate these gradients with only black-box access, zeroth-order gradient estimation techniques are adopted.

Truong et al. [63] present a similar approach, where they replace the loss function from Kullback-Leibler (KL) divergence to ℓ_1 norm loss for training the student model. In contrast to the previous attacks that generate "hard" queries that differ in predictions between the victim and clone models, Sanyal et al. [52] adopt a different strategy by generating ”diverse” queries to increase predictions belonging to different classes.

3) Backdoor Attack: Backdoor attacks [17], [32], [49], [5], [29] are training-time attacks that introduce malicious behavior into the model, making it behave like a benign model for normal inputs, while intentionally misclassifying the input to a predetermined class when the trigger appears.

The seminal work by Gu et al.[17] introduced the concept of the backdoor attack on machine learning models. Building upon this, Liu et al.[32] proposed an advanced backdooring technique that incorporates enhanced triggers and relies on fewer assumptions. However, these attacks were limited to injecting static triggers, making them susceptible to detection.

Salem et al.[49] integrated generative models to perform dynamic backdoor attacks, where the trigger is not fixed thus increasing the difficulty of detection. Nguyen and Tran[42] further extended this concept to design an input-aware attack. Most existing attacks in this domain are based on poisoning attacks [68], [40], [53], [77], which involve poisoning the training dataset. In contrast, Bagdasaryan and Shmatikov [5] propose a distinct attack target in the scenario where the learning algorithm itself is poisoned, presenting an alternative approach in this field of study.

4) Data Reconstruction Attack: Data reconstruction attacks [14], [69], [18], [48] aim to recover the target dataset with limited access to the target model, with the aid of additional knowledge possessed by the adversary.

In the realm of data reconstruction attacks, existing approaches can be broadly classified into three categories: optimization-based attacks, training-based attacks, and analysis-based attacks.

Optimization-based attacks, first introduced by Fredrikson et al.[14], represent the majority of existing reconstruction attacks. These attacks employ an iterative optimization process to reconstruct the training dataset, with the objective of obtaining a high likelihood score for the desired class. Notably, the integration of generative models by Zhang et al.[75] has contributed to improving the quality of reconstruction. Building on this line of research, several studies have explored diverse architectural choices [10], [64] and loss functions [60] to further enhance reconstruction performance.

Conversely, training-based attacks [69] regard the target model as an encoder and train a corresponding decoder network to reconstruct inputs based on the model's outputs. Recently, Haim et al. [18] presented a theoretical demonstration that, under specific assumptions, the training data can be completely recovered, leading to a new attack approach.