Please Tell Me More: Privacy Impact of Explainability through the Lens of Membership Inference Attack

Han Liu, Yuhao Wu, Zhiyuan Yu, Ning Zhang
Computer Security and Privacy Laboratory, Washington University in St. Louis, MO, USA

***Abstract*—Explainability is increasingly recognized as an enabling technology for the broader adoption of machine learning (ML), particularly for safety-critical applications. This has given rise to explainable ML, which seeks to enhance the explainability of neural networks through the use of explanators. Yet, the pursuit for better explainability inadvertently leads to increased security and privacy risks. While there has been considerable research into the security risks of explainable ML, its potential privacy risks remain under-explored.**

To bridge this gap, we present a systematic study of privacy risks in explainable ML through the lens of membership inference. Building on the observation that, besides the accuracy of the model, robustness also exhibits observable differences among member samples and non-member samples, we develop a new membership inference attack. This attack extracts additional membership features from changes in model confidence under different levels of perturbations guided by the importance highlighted by the attribution maps in the explanators. Intuitively, perturbing important features generally results in a bigger loss in confidence for member samples. Using the member-non-member differences in both model performance and robustness, an attack model is trained to distinguish the membership. We evaluated our approach with seven popular explanators across various benchmark models and datasets. Our attack demonstrates there is non-trivial privacy leakage in current explainable ML methods. Furthermore, such leakage issue persists even if the attacker lacks the knowledge of training datasets or target model architectures. Lastly, we also found existing model and output-based defense mechanisms are not effective in mitigating this new attack.

1. Introduction

The field of machine learning has advanced at a remarkable pace in the past few years, leading to the deployment of sophisticated neural networks for numerous real-world tasks. For example, ChatGPT [1] and Segment Anything [2] have revolutionized natural language processing and image segmentation, while AlphaGo highlights the power of reinforcement learning [3]. However, as these powerful AI systems are increasingly deployed to support essential societal functions, there exists a pressing need for explainability of the technology, especially for safety-critical applications. To enhance the explainability, explainable machine learning is developed to shed light onto the inner workings of neural networks [4], [5], [6], [7]. Due to its unique ability to facilitate trust establishment between users and machine [8], explainable ML has been employed in various critical domains, such as computer security [9], [10], [11], medical diagnosis [12], [13], and model debugging [14].

Security and Privacy of Explainable ML: Despite being widely used in critical domains, explainable ML has been shown to suffer from security and privacy problems. In particular, a line of studies [15], [16], [17], [18], [19] reveal that current explainable methods are susceptible to adversarial examples, capable of deceiving both target neural networks and their associated explanators. However, the privacy risks of explainable ML have received less attention so far [20], [21]. The most closely related work is Shokri et al. [20]. It examined the privacy leakage risk of explainable ML using membership inference [22], which suggested limited privacy leakage compared to model performance. However, the member and non-member differences could also exist in the model robustness. Given the growing importance of explainability in machine learning, this work aims to systematically re-evaluate the interaction between explainability and privacy using both model performance and robustness as attack vectors.

Our Approach: In this paper, we present a systematic study of the privacy risks of explainable ML toward membership inference. Our exploration focuses on one of the most prevalent explanation methods, *i.e.*, attribution-based methods, and the typical ML task, *i.e.*, classification, under the more challenging black-box scenarios. There are three research questions: *Q1. Does explainable ML exacerbate privacy leakage, and if so, what's the concrete attack vector?* To this end, we designed a type of membership inference attack that attempts to extract the membership information from not only the model performance (loss) but also the model robustness (changes in confidence due to perturbations). *Q2. What are the contributing factors leading to privacy leakage?* To answer this, we use the newly designed membership interference as the basis, and examine the impact of different levels of model overfitting and explanation quality. *Q3. How do different levels of attack knowledge on data distribution and models impact privacy leakage?* To tackle this, we examine the privacy leakage levels by varying the attack knowledge of model architectures, the degree of overfitting, and different data distributions in shadow models. There are three technical challenges for this new attack.

First, the high-dimensional nature of images poses a challenge in determining what to perturb in the sample. To address this, we utilize two perturbation strategies: perturbing the most important pixels first (MoRF) and the least important pixels first (LeRF) as indicated by the attribution maps from the explanator. Adjusting the perturbation levels under these strategies allows the attack model to observe the difference in confidence drops between member samples and non-member samples, while increasingly/decreasingly important features are perturbed.

Second, the perturbation operation may introduce distribution shifts and adversarial artifacts, complicating the determination of whether the degradation in prediction performance results from the perturbation of important features highlighted by the explanator. To tackle this challenge, we propose a new perturbation operation, which approximates the values of perturbed pixels using the weighted mean of their neighbors to mitigate the distribution shift. Additionally, we introduce Total Variation (TV) as a regularization method to tackle adversarial artifacts.

Third, prediction outcomes may be too noisy to reflect the membership status. Therefore, we employ hypothesis testing to select critical trajectories, which can rigorously quantify the statistical significance of the relationship between perturbation trajectories and membership status.

Evaluation and Findings: We extensively evaluate our attack with different explanation methods, benchmark datasets, and model architectures. In particular, we investigate four types of explanations, spanning seven state-of-the-art explanation methods, including SmoothGrad [6], VarGrad [7], IG [23], Grad-CAM [4], Grad-CAM++ [5], LIME [24], and SHAP [25]. The results show significant and pervasive privacy leaks in these ML explanation methods. It is worth noting that even when adversaries lack knowledge of the architecture and training datasets of the target models, privacy leakage remains severe. Further analysis indicates a possible trade-off between the quality of explanations and the risk of privacy leakage. Notably, methods providing superior explanation quality may also present an increased potential for privacy breaches. Furthermore, we empirically assess two widely utilized defensive strategies, yet none of the current methods provide an effective shield against the leakage stemming from explainability.

Contributions: Our contributions are outlined as follows:

- We present a systematic study to assess the privacy risks of explainable ML through the lens of membership inference. We introduce a novel and generalizable method for extracting membership features using both model performance and model robustness.
- We conduct comprehensive experiments that expose significant privacy risks that span a variety of explanators, datasets, and model architectures.
- We provide in-depth analyses of the privacy leakage factors caused by explainable ML. Additionally, we investigate the influence of each design component and threat model on the efficacy of our attacks.

2. Background

2.1. Membership Inference Attacks

Membership inference attacks pose significant threats to privacy, as they can reveal confidential information used in the development of machine learning models. In such attacks, an adversary aims to determine if a specific data point was used to train an ML model. Formally, given a sample x, a trained model $\mathcal{M}$, a membership inference attack can be defined as:

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

where $\mathcal{A}$ denotes the attack model, most attack models are essentially binary classifiers [26], [27], [28], [29], and can be constructed in various ways, depending on the underlying assumptions. If the data sample x was used to train the model $\mathcal{M}$, the attack model outputs 1 (*i.e.*, member); otherwise, it outputs 0 (*i.e.*, non-member).

Owing to the practical threat models, the majority of membership inference attacks focus on black-box settings [22], [26], [27], [30], [31], [32], where the adversary only has access to the target model's posterior output. A widely used approach involves training shadow models to mimic the behavior of target models. These shadow models are then employed to generate features for training a neural network-based attack model [22], [26], [27], [33]. Additionally, several studies have introduced metric-based attacks that determine a global threshold to differentiate membership status without requiring attack models, relying on entropy loss [34] and its variants [30].

More recently, Liu et al. [26] employed membership information from the model's full training process, where they found that member samples exhibited unique loss trajectories. By performing model distillation using auxiliary datasets and assessing losses on intermediate models to obtain loss trajectories, they effectively facilitated membership inference. Carlini et al. [32] introduced a novel line of research examining per-sample hardness of privacy leakage, called Likelihood Ratio Attack (LiRA). They trained a large number (e.g., several hundred) of shadow models per target sample and inferred the confidence distribution of these samples. Subsequently, they executed a parametric likelihood-ratio test to ascertain membership.

2.2. Explainable Machine Learning

As machine learning systems are increasingly deployed in critical domains [35], [36], [37], explainable ML has gained widespread popularity for improving the understanding of decision-making processes inherent in machine learning models and delineating the factors or processes that guide these decisions. The concept of explainable ML is closely tied to interpretable ML; however, these two concepts are inherently different [38]. Interpretable ML focuses on designing intrinsically interpretable models, whereas explainable ML aims to provide post hoc explanations for existing black-box models [9], [39].

Given that explainable machine learning necessitates no alterations to the underlying model architecture, it proves particularly beneficial for intricate black-box models, such as neural networks, which often attain superior accuracy levels [9], [40]. In this paper, we primarily consider explainable ML, specifically attribution-based explainable methods that have become the most studied in recent years [4], [6], [41], [42], [43]. These methods aim to generate pixel attribution maps identifying the contributions of each feature (*i.e.*, pixels) in the input towards a specific model prediction. The four leading attribution-based explanation methods are backpropagation, representation, perturbation, and approximation-based explanations [35], [44], [45].

Backpropagation-based Explanations. These methods derive feature attribution maps by computing the gradient (or its variants) of the model prediction with respect to a given input [41], [42], [46]. The underlying principle is that a larger gradient signifies greater relevance of the feature to the prediction. Specifically, the attribution maps of input x are derived based on

$$g^c(x) = \frac{\partial f_c(x)}{\partial x}, \tag{2}$$

where $f_c(x)$ represents the model prediction for a given input x and a given class c. Integrated Gradients (IG) [23] is designed to meet three desirable axioms of explainability: sensitivity, implementation invariance, and completeness. It generates attribution maps through the calculation of the path integral of the gradients along a linear path from a baseline (*e.g.*, a zero input) to the input x. However, directly produced attribution maps can be noisy and may not effectively reflect meaningful information. To resolve this, SmoothGrad [6] is proposed to visually sharpen gradient-based attribution maps by adding Gaussian noise to original samples and then calculating the average attribution maps among all samples to obtain the explanation. VarGrad is proposed [7], which follows a similar process of adding Gaussian noise to original samples but calculates the variance of attribution maps instead, which could capture higher-order partial derivative information of the model prediction [47].

Representation-guided Explanations. These methods employ feature maps at intermediate layers of models to produce attribution maps [4], [5], [48]. The core idea is that higher layers of a CNN capture higher-level semantics and detailed spatial information. Specifically, they calculate the attribution maps based on

$$g^c(x) = \sum_k \alpha_k^c A^k, \tag{3}$$

where A^k represents the k-th feature map, and α_k^c represents the importance of feature map k to class c. To calculate α_k^c, Grad-CAM [4] proposes employing the gradients flowing into the feature maps to calculate weights. Grad-CAM++ [5] further improves explanation quality by refining the way gradients are weighted, considering the pixel-wise contribution of each gradient when calculating feature map weights.

Perturbation-based Explanation. These methods measure the contribution of each feature by observing the changes in the prediction score when the feature is perturbed. A notable method is SHapley Additive exPlanations (SHAP) [25]. SHAP calculates the contribution of a feature to the prediction as the average difference in the model's output when that feature is included and excluded, considering all possible feature combinations:

$$g_i^c(x) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup i) - f(S)], \tag{4}$$

where N is the set of all features, S is a subset of N that does not include feature i, $f(S)$ is the prediction of the model when only the features in set S are used, and $|S|$ denotes the size of set S. The exact computation of SHAP values is computationally intensive, thus the paper further proposes various methods to approximate them, such as kernel SHAP, Deep SHAP, etc.

Approximation-based Explanation. These methods derives explanations by approximating the predictions of original models in the vicinity of a given input by an interpretable model. One representative approach is Local Interpretable Model-agnostic Explanations (LIME) [24]. It obtains the optimum interpretable models by optimizing a loss function over a set of potentially interpretable models (e.g., linear models, decision trees, etc.) Specifically, it solves the following optimization problems:

$$g^c(x) = \text{argmin}_{i \in I} \mathcal{L}(f, i, \pi_x) + \Omega(i), \tag{5}$$

where $\mathcal{L}$ measures the local fidelity of the function f in the proximity of x, as approximated by the interpretable model i. This locality is measured by π_x. Meanwhile, $\Omega(i)$ is utilized to constrain the complexity of the interpretable model i, thereby promoting simplicity and interpretability.

3. Threat Model

Adversary's Goal. Similar to membership interference attacks on classical machine learning models, the adversary's goal in launching a membership inference attack on explainable machine learning is to infer whether the target sample is used to train the original model. However, different from the membership interference attack on the classical model, explainable machine learning offers additional access to both the original model and the explanator that is associated with the original model. Membership inference attack in several recent works [26], [29], [32], [49] has demonstrated significant privacy risks. Given that explainable machine learning is increasingly adopted in privacy-sensitive scenarios such as medical domains [50], financial domains [51], and social media domains [52], it is important to investigate whether these explainable techniques would inadvertently help the adversary gain an additional advantage in privacy breaches.

Adversary's Knowledge and Capabilities. We assume a black-box setting in our attacks: for the target model, the adversary only has access to the posterior output without

TABLE 1: Performance comparison of existing membership inference attacks using explanations and model outputs against baselines using model outputs.

Attack Method	Explanation Method	TPR at 0.1% FPR			
		CIFAR-100	CIFAR-10	CINIC-10	GTSRB
Shokri et al. [22]	None	0.5%	0.2%	0.5%	0.1%
Shokri et al. (expl.) [20]	SmoothGrad	0.4%	0.2%	0.2%	0.3%
	VarGrad	0.3%	0.2%	0.2%	0.1%
	IG	0.4%	0.3%	0.3%	0.1%
	GradCAM	0.6%	0.3%	0.2%	0.1%
	GradCAM++	0.6%	0.2%	0.2%	0.3%
	SHAP	0.4%	0.2%	0.3%	0.2%
	LIME	0.4%	0.1%	0.3%	0.2%

knowing parameters; for explanators, the adversary only has access to attribution maps and is unaware of method details and parameters. Moreover, we assume the adversary knows the architecture of the target model and has an auxiliary dataset that comes from the same distribution as the training dataset of the target models, which is a common setting in most existing works [27], [29], [30], [31], [53], [54], [55].

4. Preliminary Exploration and Motivation

Despite the critical importance of the issue, privacy leakage in explainable ML remains a largely unexplored domain. Shokri et al. [20] made a first effort to delve into the inherent vulnerabilities of explainable ML. Their study primarily relied on explanations as the sole source of information, revealing that the variance of explanations can disclose membership status. The core observation is that high explanation variance indicates proximity to a decision boundary, which is more common in non-member samples. However, the field of membership inference continues to evolve, leading to the establishment of more structured adversary knowledge and standardized evaluation settings. Moreover, the target explanations presented by existing works do not include all popular types of attribution-based methods. In this work, the same techniques are studied under these new conditions, including the newly accessible adversary information (*i.e.*, target model output), evaluation settings, and more extensive explanations. In particular, experiments are carried out across four popular types of explanations, inconsistent with the evaluation settings of the latest membership inference research [26], [29], [32]. Our method incorporates shadow model training techniques and consequently trains an attack model that employs both the explanation and loss as input. To evaluate whether the explanation reveals additional leakage, we incorporate a representative membership inference attack [22], which relies solely on target model outputs as information sources, serving as a baseline measure.

Experimental Setup. Our target model employs the ResNet-18 architecture [56]. For the attack model, we use a Multilayer Perceptron (MLP) architecture with fully connected layers of dimensions [r, 2048, 1024, 512, 256, 64, 2], where r is the combined dimension of the explanation vector and loss. Four benchmark datasets were used: CIFAR-10 [57], CIFAR-100 [57], CINIC-10 [58], and GTSRB [59]. We implemented baselines in the same dataset and model setting. We assess our results using TPR at a fixed 0.1% FPR, following [26], [32]. Detailed settings are provided in section 6.1. The effectiveness of membership inference by using explanations only is evaluated in Appendix A.

Results and Analysis. As depicted in Table 1, the attack leveraging both the target model output and explanations yields results similar to those attacks using model output only, which are consistent with findings of [20]. The reason is that the explanation is noisy [6] and its variance fluctuates wildly between individual images, thus making it challenging to infer membership based on explanation variance. Furthermore, explanations often contain substantial redundant information that does not pertain to its membership status. For instance, different classes may have varying distributions, providing useful classification information, but they do not offer useful membership status information.

Motivation. From the results of existing methods, one may infer that current explanation methods exhibit limited privacy leakage relative to model performance. However, in this paper, we found that explanation can lead to significant privacy leakage in the context of membership inference. To this end, we aim to address the following questions throughout our research:

Q1. Does explainable ML exacerbate privacy leakage, and if so, what's the concrete attack vector?
Q2. What are the contributing factors leading to privacy leakage?
Q3. How do different levels of attack knowledge on data distribution and model impact privacy leakage?

To answer these questions, we will revisit the current state-of-the-art explanations systematically, identify universal vulnerabilities in different types of explanations irrespective of their underlying mechanisms, and then design advanced attacks based on these vulnerabilities to fully reveal the privacy leakage inherent in the explanations.

5. Attack Methodology

5.1. Problem Formulation

Formally, given a sample x, a target model $\mathcal{M}$, and a coupled explanator $\mathcal{G}$, a membership inference attack can be defined as

$$\mathcal{A} : x, \mathcal{M}, \mathcal{G} \rightarrow \{0, 1\}, \tag{6}$$

where 1 means x is a member of $\mathcal{M}$, and 0 means x is a non-member. Different from traditional membership inference which only utilizes the posterior outputs from the target model $\mathcal{M}$ (*e.g.*, loss) [26], [31], [34], our attacks require considering the additional outputs from the explanator $\mathcal{G}$, which are generally attribution maps. This introduces challenges on how to leverage this additional information to enhance the attack performance. As demonstrated in the prior section, sole reliance on explanation variance fails to generate meaningful features related to membership status,

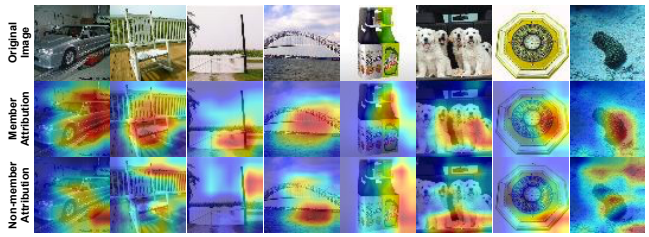


Figure 1: Attribution maps of member samples and non-member samples. The first row shows the original images, the second row and third row show the attribution maps when the original images serve as member samples and non-member samples, respectively.

necessitating a new attack method. Moreover, considering the distinct technical mechanisms underlying various explanation methods, as depicted in Section 2, it becomes vital to conduct an analysis from a more general perspective in order to identify common vulnerabilities. Therefore, we initiate our investigation with a focus on the generalized properties of explainable machine learning.

5.2. Design Intuition

ML models are designed to fit training data, allowing them to make predictions on unseen data. However, this can lead to generalization gaps between training data and testing data [60], which have been identified as a major contributing factor to the success of membership inference attacks [22], [27]. These generalization gaps lead to differences in model performance metrics, such as loss differences, between member and non-member samples, which serve as the primary features used by most previous works [26], [30], [31], [32], [34] to infer membership status.

However, a less explored aspect is how these generalization gaps affect the attribution maps of explanators. To investigate this, we use the same set of samples in two distinct scenarios. In the first scenario, the set of samples is used to train a model, acting as member samples; in the second scenario, the set of samples does not participate in the training process, serving as non-member samples. We generate attribution maps for these samples in both trained models. We repeat the above experiments for different datasets, explanators, and model architectures, and we leverage Structural Similarity Index Measure (SSIM) [61] to measure the similarity between members and non-members. Specifically, on the CIFAR-100 dataset, we observed the following SSIM values for attribution maps between members and non-members: SmoothGrad (0.56), VarGrad (0.67), IG (0.12), GradCAM (0.73), GradCAM++ (0.52), SHAP (0.72), and LIME (0.34). Some examples are shown in Figure 1. Our analysis indicates that member and non-member samples generate distinct explanation patterns, potentially stemming from inherent robustness disparities. Notably, members tend to focus on key semantic features crucial for classification, making their classification more susceptible to changes if these features are altered. Conversely, non-members either

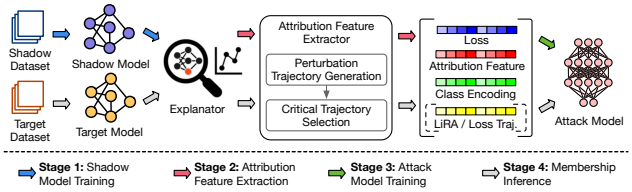


Figure 2: Overview of our attack pipeline.

do not concentrate on or only partially engage with these key features. Consequently, modifications to these features tend to have a comparatively minor impact on the classification outcomes of non-member samples.

In this work, we demonstrate that the robustness differences between members and non-members may expose a new attack surface. An adversary could utilize the explanations as guidance to perturb the image and make predictions again. Member samples and non-member samples are expected to exhibit different confidence score changes under such perturbations. In other words, perturbing the informative features highlighted by explanators should result in a more significant decrease in confidence scores in members compared to non-members. Therefore, the adversary can leverage the different confidence changes under perturbations to determine whether a sample is a member or not.

5.3. Attack Overview

The general attack pipeline of our attack is given in Figure 2. It consists of four stages: shadow model training, attribution feature extraction, attack model training, and membership inference.

Shadow Model Training. To mimic the behavior of the target model $\mathcal{M}$, the adversary trains a shadow model $\mathcal{M}^S$. As mentioned earlier, the adversary has an auxiliary dataset $\mathcal{D}^a$ drawn from the same distribution as the training dataset of the target model. The adversary splits the auxiliary dataset into two disjoint subsets, and one subset is used as the shadow dataset $\mathcal{D}^a_{shadow}$ to train the shadow model.

Attribution Feature Extraction. Given a target sample x, the adversary queries the trained shadow model and uses the explanator $\mathcal{G}$ to generate attribution maps to explain the decisions. Then, different levels of perturbations are applied to the sample x guided by the attribution maps, and the prediction changes are measured. Such prediction changes with respect to different perturbation levels form perturbation trajectories, and we select the most informative trajectories through hypothesis testing. We elaborate on the whole process in Section 5.4.

Attack Model Training. Utilizing the attribution features, the adversary aggregates them with the loss computed through querying the shadow model $\mathcal{M}^S$ and the one-hot encoding of the classes to construct membership features. We could extend the applicability of our methods to other advanced membership inference frameworks by simply integrating their attack features. Subsequently, a Multi-Layer

Perceptron (MLP) network is trained to serve as the attack model, inferring membership based on the integrated membership features. We provide more details in Section 5.5.

Membership Inference. In the final step, the adversary inputs various features into the trained attack model, including those derived from the target model (*e.g.*, losses) and attribution features acquired from the explanator, one-hot encoding of the classes, and potentially additional attack features. Subsequently, the attack model generates a binary membership status of the target sample.

5.4. Attribution Feature Extraction

As previously mentioned, an adversary can exploit explanations as guidance to perturb images and use the prediction score changes as features to distinguish between members and non-members. However, generating such perturbations has several technical challenges:

C1. Perturbation Strategy Determination: The high-dimensional nature of images poses a challenge in determining what elements to perturb in the sample. This complexity underscores the necessity to devise optimal perturbation strategies, which allow for accurate differentiation between members and non-members while reducing the number of model queries.

C2. Distribution Shifts and Adversarial Artifacts: Perturbation operations have been shown to induce distribution shifts [62] and potentially introduce adversarial artifacts [14], [63]. As a result, it becomes difficult to ascertain whether the degradation in prediction performance originates from the distribution shift (or adversarial artifacts) or the perturbation of informative features.

C3. Attribution Feature Selection: Given the influences of distribution shifts and adversarial artifacts, the prediction outcomes may be too noisy to accurately reflect membership status. Therefore, it is crucial to select the most informative attribution features generated by perturbations to effectively infer membership status.

Perturbation Trajectory Generation. To address challenges C1 and C2, we propose a novel perturbation trajectory generation technique. This method involves designing optimal perturbation strategies to efficiently extract membership features while minimizing the number of queries. Furthermore, we develop an enhanced perturbation operation to mitigate distribution shifts and adversarial artifacts. In order to design an effective perturbation strategy, we draw upon practices in existing explanation evaluation methods [64], [65]. This evaluation is crucial for building trust in the accuracy of the explanations and ensuring that they truly reflect the important features of decision-making processes. It could also facilitate the benchmarking of various explanation methods [62]. Specifically, we adopt the two widely-used perturbation strategies: Most Relevant First (MoRF) and Least Relevant First (LeRF) [64], [66]. MoRF perturbs the most relevant pixels first; when applied to our scenario, the model confidence score should decrease more rapidly for member samples that focus on key semantic features. On the other hand, LeRF perturbs the least relevant pixels first, so the model output should change more slowly for member samples in this case. We use different perturbation percentages (*i.e.*, percentages of pixels perturbed) to iteratively perturb the images and record the confidence drop at each percentage. As a result, we obtain a confidence score drop trajectory with respect to the perturbation percentages.

Before elaborating on our solutions to address C2, we first explain why distribution shifts and adversarial artifacts may occur. Generally, the perturbation operation uses a fixed value (*e.g.*, zero value) to replace the corresponding pixels, as in many explanation evaluation methods [67], [68]. However, such an abrupt introduction of artifacts can change the original training distribution, making it unclear whether the degradation in model performance comes from the distribution shift or because the truly informative features are removed. To counteract this distribution shift, we adopt the Noisy Linear Imputation as our perturbation operator proposed in Remove and Debias (ROAD) [65]. Specifically, it approximates the values of perturbed pixels by the weighted mean of their neighbors. When multiple pixels are perturbed, it sets an equation system for each pixel. Then, they plug in the values of known pixels and consider the perturbed pixels as unknown variables. The linear equation system can then be solved efficiently. Since every perturbed pixel is highly correlated with existing pixels, the distribution shift caused by perturbation can be mitigated.

Perturbation operations may also introduce adversarial artifacts [14], [63], where small changes in input data can lead to unpredictable effects on the output. Since adversarial perturbations often come in the format of unnatural, unconstructed noise [43], [69], we aim to regularize our perturbation to have a more natural and smooth pattern to mitigate the adversarial artifacts. To this end, we adjust the priority of choosing pixels as

$$I_{ij} = g_{ij} - \alpha \cdot TV(m_{ij}), \tag{7}$$

where g_{ij} is the attribution for the position (i, j) in the original image $m_{i,j}$, and TV (Total Variation) is used to limit the pixel difference, which is calculated as

$$TV(m_{ij}) = \sum_{m=-1}^{1} \sum_{n=-1}^{1} |m_{i+m,j+n} - m_{i,j}|. \tag{8}$$

Incorporating the TV term in the priority calculation promotes the generation of smoother and more natural patterns. This regularization can reduce the impact of adversarial artifacts, making it easier to differentiate between changes caused by the perturbation of informative features and those arising from adversarial noise.

Critical Trajectory Selection. To address C3, we propose critical trajectory selection to extract the most informative features that differentiate members and non-members. Since different images may exhibit statistical variances in their confidence drops within each perturbation percentage, we propose employing hypothesis testing to rigorously quantify the statistical significance of the relationship between confidence drops at varying perturbation percentages and

membership statuses. Using hypothesis testing can help us confidently select features that are truly informative, as opposed to those that appear important due to random chance.

We first formulate the null hypothesis (H_0) and the alternative hypothesis (H_1) for each feature (*i.e.*, confidence score drop at certain perturbation percentage) with respect to its impact on the target variable (*i.e.*, membership status). Suppose Q^M represents the features of member samples, and Q^{NM} represents the features of non-member samples, $\mathcal{A}$ represents the membership status, then for the i-th feature, we define the hypotheses as

$$\mathcal{H}_{0i} : P_r(\mathcal{A} = 1|D_i \in Q^M) = P_r(\mathcal{A} = 1|D_i \in Q^{NM}), \tag{9}$$

$$\mathcal{H}_{1i} : P_r(\mathcal{A} = 1|D_i \in Q^M) \neq P_r(\mathcal{A} = 1|D_i \in Q^{NM}). \tag{10}$$

We observe that the confidence drops of different samples within the same perturbation percentage can be treated as a normal distribution according to a Shapiro-Wilk Test [70], and they have different variances. Therefore, we apply Welch's t-test [71] to test the hypothesis. Specifically, the confidence drops for member samples and non-member samples are denoted as $m_1, \ldots, m_{k_m}$ and $n_1, \ldots, n_{k_n}$, respectively, so that the t statistic can be calculated by

$$t = \frac{\overline{m} - \overline{n}}{\sqrt{\frac{s_m^2}{k_m} + \frac{s_n^2}{k_n}}}. \tag{11}$$

where $\overline{m}$ and $\overline{n}$ are the means of the confidence drops for members and non-members, s_m^2 and s_n^2 are the corresponding unbiased estimators of the population variance. The degree of freedom (ν) is approximated using Welch-Satterthwaite equation [72] as

$$\nu \approx \frac{\left(\frac{s_m^2}{k_m} + \frac{s_n^2}{k_n}\right)^2}{\frac{(s_m^2/k_m)^2}{k_m - 1} + \frac{(s_n^2/k_n)^2}{k_n - 1}}. \tag{12}$$

We then obtain the corresponding p-value as

$$p = Pr(T_\nu > |t|), \tag{13}$$

where T_ν is a random variable following the t-distribution with ν degrees of freedom. Since the p-value is the probability of observing the test statistic as extreme as the one calculated, assuming the null hypothesis is true, a lower p-value indicates a higher probability of rejecting the null hypothesis (*i.e.*, the feature has a significant impact on the target variable). To this end, we rank the features based on their p-values, selecting a subset of features with the lowest p-values for attack model training.

To further improve the attack performance, we perform membership inference attacks on the augmented versions of the example since models are typically trained to minimize their loss not only on the original training example but also on the augmented versions of the example [32]. To achieve this goal, we query the original models multiple times to obtain the loss of augmented examples as features for membership inference. Explanators could also be applied to these augmented examples to obtain attribution maps, which can be used to generate additional perturbation trajectories.

5.5. Attack Model Design

The attack model utilizes the extracted features to determine the binary membership status. We have three different types of features: MoRF-based perturbation trajectories, LoRF-based perturbation trajectories, and losses. We also incorporate the one-hot encoding of classes as a feature, given that different classes may exhibit unique perturbation trajectories. Consequently, the input to the attack model comprises the concatenation of the aforementioned features:

$$\hat{a} = \mathcal{J}_{\text{MoRF}}(g) \oplus \mathcal{J}_{\text{LeRF}}(g) \oplus \mathcal{L}(x) \oplus \mathcal{O}(c), \tag{14}$$

where $\mathcal{J}_{\text{MoRF}}(g)$ and $\mathcal{J}_{\text{LeRF}}(g)$ represents the perturbation trajectories generated on explanation g in the MoRF and LeRF settings, respectively. $\mathcal{L}(x)$ represents the loss of the model, and $\mathcal{O}(c)$ represents the one-hot encoding of classes c of the given sample x, and $\hat{a}$ is the input to the attack model, and the corresponding label is 1 if x is used for training the shadow model and 0 otherwise. In Section 6.2, we demonstrate how our attacks can be readily extended to the scenarios of other advanced membership inference attacks, simply by concatenating their attack features (for instance, trajectory loss in [26] and LiRA confidence in [32]) into $\hat{a}$, which can significantly enhance the attack performance in their respective contexts. For the attack model, we use an MLP architecture with fully connected layers of dimensions [r, 1024, 512, 128, 32, 2] with ReLU activation function, where r is the dimension of $\hat{a}$. A Softmax function is applied to the last FC layer to obtain the final output.

6. Experiments

In this section, we design our experiments to answer the three research questions in Section 4 by conducting an empirical study of our methods on a variety of explanation methods, datasets, and models. Specifically, we answer *Q1* in Section 6.2, *Q2* in Section 6.3, and *Q3* in Section 6.4.

6.1. Experimental Setup

Datasets. We use four benchmark datasets, CIFAR-10 [57], CIFAR-100 [57], CINIC-10 [58], and GTSRB [59], that are common in membership inference attack studies for our experiments. Following existing works [27], [28], [29], [73], each dataset is split into four equal subsets: $\mathcal{D}_{target}^{train}$, $\mathcal{D}_{target}^{test}$, $\mathcal{D}_{shadow}^{train}$, and $\mathcal{D}_{shadow}^{test}$. Data samples in $\mathcal{D}_{target}^{train}$ are used to train a target model $\mathcal{M}$ and are considered as members of $\mathcal{M}$, while data samples in $\mathcal{D}_{target}^{test}$ are considered as non-members. Similarly, data samples in $\mathcal{D}_{shadow}^{train}$ are used to train a shadow model $\mathcal{M}^S$ and are treated as members and data samples in $\mathcal{D}_{shadow}^{test}$ as non-members. $\mathcal{D}_{shadow}^{train}$ and $\mathcal{D}_{shadow}^{test}$ are used to create membership features for training the attack model $\mathcal{A}$.

Model Architectures. We employ a widely used ResNet-18 [56] to serve as target and shadow models in our quantitative evaluation. In our ablation study, we evaluate our attack performance under the settings of different model

architectures: ResNet-56 [56], VGG-16 [74], and DenseNet-161 [75]. To reduce model overfitting, we train models with standard techniques, including weight decay [76] and train-time augmentations [77]. Detailed data split, training configurations, and model accuracy are provided in Appendix B.

Target Explanators. We evaluate seven representative explanators: SmoothGrad [6] VarGrad [7], and IG [23] for the backpropagation-based explanation; Grad-CAM [4] and Grad-CAM++ [5] for the representation-guided explanation; SHAP [25] for the perturbation-based explanation, and LIME [24] for the approximation-based explanation. For each sample, we generate attribution maps corresponding to the category with the highest score.

Evaluation Metrics. We employ the following metrics, in line with recent state-of-the-art works [26], [32]: (1) *Full Log-scale receiver operating characteristic (ROC)*, a widely used ROC curve, reported on a logarithmic scale to emphasize low false-positive rates; (2) *TPR at Low FPR*, measuring attack performance at a fixed FPR (*e.g.*, 0.1%), served as a straightforward comparison metric. (3) *Balanced Accuracy and area under the ROC curve (AUC)*, widely-used average case metrics in existing membership inference attacks [22], [27], [29], [31]. The balanced accuracy measures attack prediction accuracy on a balanced dataset of members and non-members. We include this metric for completeness as [26], [32], though it may not be the most suitable.

Comparison Baselines. We take Shokri et al. [20] (*aka*, *Shokri et al. (expl.)*) as the benchmark for comparison. Additionally, to systematically answer *Q1*, we compare the proposed attack with state-of-the-art membership inference attacks [22], [27], [31], [34]. These comparisons adhere to the standard adversarial settings discussed in Section 3, utilizing only the target model's output as information sources. In addition to these standard threat models, we also consider two additional settings that were studied in previous works [26], [32]. Across all these scenarios, our attacks are implemented under the same adversary settings as baseline methods but incorporate additional explanation knowledge. We maintain consistency in our evaluation by utilizing the same set of shadow and target models. This approach enables us to quantitatively measure the additional privacy leakage introduced by explanators.

6.2. Quantitative Evaluation

In this section, we aim to answer *Q1* by presenting extensive attack results across various explanators and datasets, and comparing them to attack results of state-of-the-art membership inference methods that rely on model output. Given the variability in threat models followed by different membership inferences, we assess them individually. Our evaluation encompasses three distinct settings. The first aligns with the standard membership inference settings adopted by most methods as discussed in Section 3. The second setting was adapted by [26], it is assumed that the adversary possesses supplementary large datasets in addition to the existing auxiliary datasets. This additional dataset can be used for model distillation. The third one was adapted by [32], it is assumed that the adversary can train a large number (e.g., several hundred) of shadow models. These models are used to generate customized confidence scores that consider the per-sample hardness.

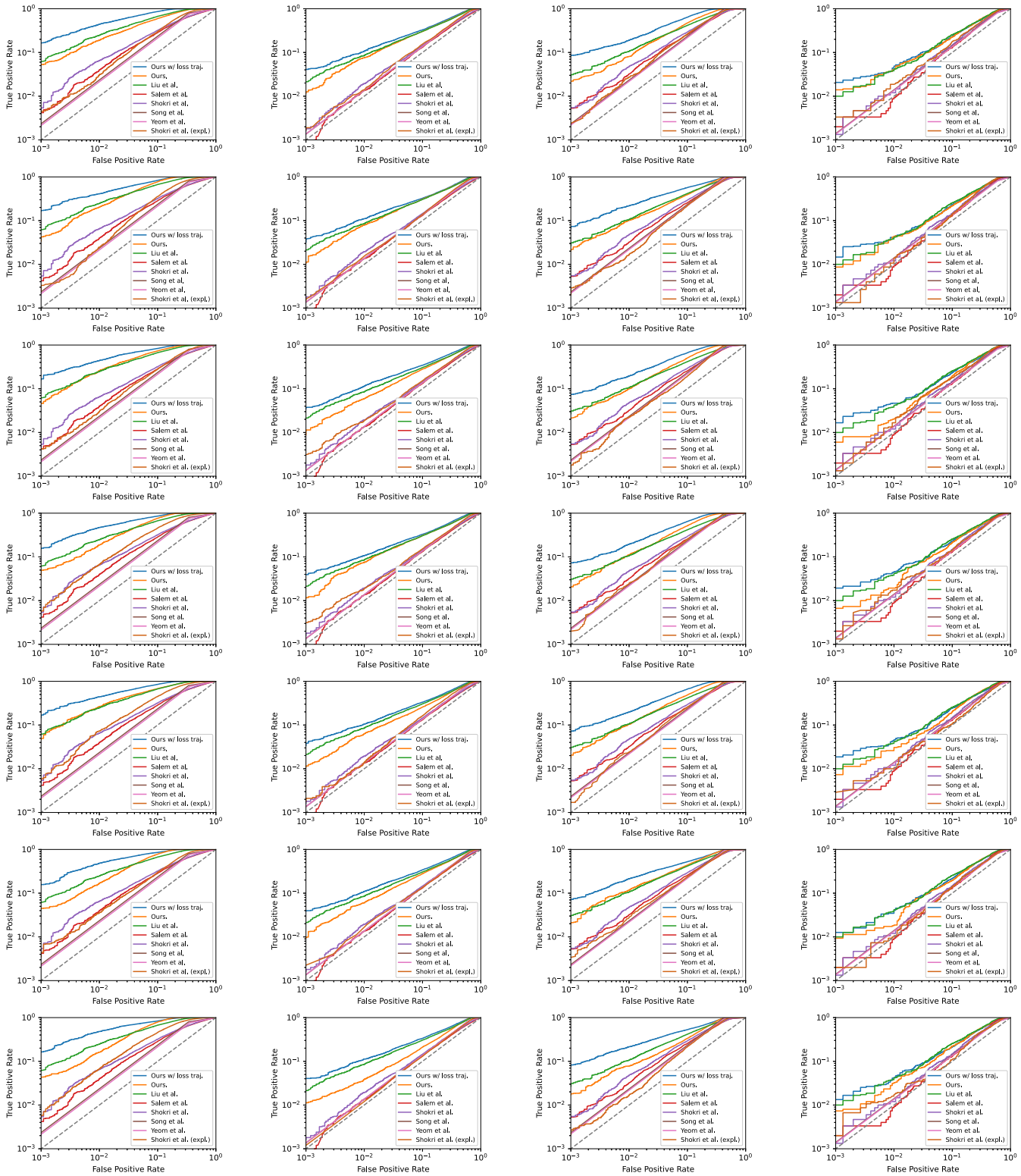
Evaluation of Standard Settings. In the standard settings, we consider the baseline of Shokri et al. (expl.) [20], as well as baselines that are solely reliant on model outputs: Yeom et al. [34], Song et al. [31], Salem et al. [27] and Shokri et al. [22]. Figure 3 illustrates the ROC curve of our evaluation results. In addition, the numerical results are provided in the upper half of Table 2. We also apply augmentations to baselines, the results are given in Table 14. Our attacks outperform the baselines [22], [27], [31], [34], particularly in the low-FPR regime. For example, when employing Grad-CAM methods, our attacks achieve a 5.0% TPR at 0.1% FPR on the CIFAR-100 dataset. In comparison to the 0.5% TPR in the baselines, such a result nearly increases the privacy leakage tenfold. Also, our attack consistently exhibits better performance in terms of balanced accuracy and AUC. Furthermore, our method outperforms Shokri et al. (expl.), even though the latter incorporates the same type of knowledge as we do. Additionally, the extent of privacy disclosure varies across different explanators. We discuss the potential reasons in Section 6.3.

Evaluation of Settings in Liu et al. [26]. In this scenario, the assumption is that adversaries have access to large supplementary datasets. We follow the original paper to execute model distillation using these additional datasets and to extract loss trajectories. Subsequently, we combine our attack features with these loss trajectories (denoted as *Ours w/ loss traj*) to form the new attack features for training the attack model. Figure 3 and the lower half of Table 2 present a comparison of our attacks in this setting with Liu et al. Our attack performance is notably enhanced compared to the use of only loss trajectories.

Evaluation of Settings in Carlini et al. [32]. In this setting, the assumption is that the adversary is capable of training a significant number of shadow models. We follow the original LiRA attack, training 256 shadow models (128 IN models and 128 OUT models). Additionally, we explore scenarios where the adversary's ability to train a multitude of shadow models is limited, whereby we train merely 16 shadow models (8 IN models and 8 OUT models). We generate confidence scores for each sample in the IN and OUT models, and subsequently feed our attack features concatenated with confidence scores to the attack model. Since each sample requires training an attack model, to reduce attack costs, we instead train an SVM. The experimental results, presented in Table 3, demonstrate that incorporating explanations can enhance attack performance, particularly when the number of available shadow models is restricted.

6.3. Qualitative Analysis

In this section, we dive deeper into the reasons behind the severe privacy leakage in explainable ML and investigate



(a) CIFAR100 (b) CIFAR10 (c) CINIC-10 (d) GTSRB

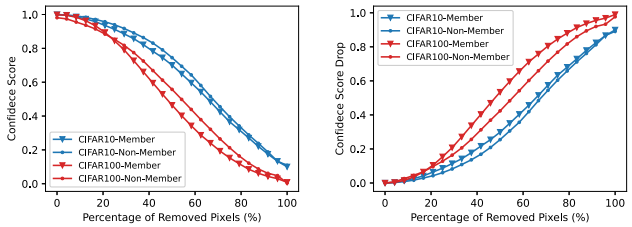
Figure 3: ROC curves for attacks on four different datasets and seven explanators (from top to bottom: SmoothGrad, VarGrad, IG, Grad-CAM, Grad-CAM++, SHAP, and LIME).

TABLE 2: Comparison of our attack with various membership inference attacks on distinct explanators and datasets.

Attack Method	Explanation Method	TPR at 0.1% FPR				Balanced Accuracy				AUC			
		CIFAR-100	CIFAR-10	CINIC-10	GTSRB	CIFAR-100	CIFAR-10	CINIC-10	GTSRB	CIFAR-100	CIFAR-10	CINIC-10	GTSRB
Yeom et al. [34]	None	0.2%	0.1%	0.2%	0.1%	0.755	0.618	0.758	0.615	0.705	0.618	0.758	0.606
Song et al. [31]		0.2%	0.1%	0.2%	0.1%	0.755	0.617	0.757	0.615	0.730	0.617	0.757	0.616
Salem et al. [27]		0.4%	0.0%	0.5%	0.2%	0.656	0.552	0.697	0.602	0.703	0.569	0.746	0.649
Shokri et al. [22]		0.5%	0.2%	0.5%	0.1%	0.661	0.565	0.703	0.600	0.718	0.591	0.761	0.639
Shokri et al. (expl.) [20]	SmoothGrad	0.4%	0.2%	0.2%	0.3%	0.741	0.607	0.750	0.500	0.799	0.642	0.782	0.679
	VarGrad	0.3%	0.2%	0.2%	0.1%	0.754	0.616	0.758	0.658	0.814	0.638	0.774	0.692
	IG	0.4%	0.3%	0.3%	0.1%	0.712	0.594	0.711	0.637	0.777	0.630	0.758	0.687
	GradCAM	0.6%	0.3%	0.2%	0.1%	0.775	0.614	0.697	0.578	0.843	0.654	0.781	0.642
	GradCAM++	0.6%	0.2%	0.2%	0.3%	0.765	0.621	0.701	0.623	0.842	0.659	0.782	0.613
	SHAP	0.4%	0.2%	0.3%	0.2%	0.751	0.607	0.714	0.616	0.798	0.618	0.760	0.638
	LIME	0.4%	0.1%	0.3%	0.2%	0.744	0.604	0.692	0.610	0.788	0.616	0.757	0.627
Ours	SmoothGrad	5.2%	1.3%	2.2%	1.4%	0.822	0.641	0.800	0.650	0.938	0.729	0.862	0.713
	VarGrad	4.3%	1.1%	2.0%	0.9%	0.815	0.639	0.801	0.607	0.940	0.727	0.833	0.700
	IG	4.5%	1.1%	2.1%	0.6%	0.808	0.638	0.815	0.595	0.932	0.714	0.878	0.693
	GradCAM	5.0%	1.2%	2.1%	0.7%	0.783	0.647	0.814	0.615	0.938	0.722	0.882	0.682
	GradCAM++	4.9%	1.2%	2.1%	0.7%	0.799	0.652	0.800	0.615	0.938	0.722	0.862	0.682
	SHAP	4.5%	1.0%	2.1%	0.9%	0.857	0.644	0.776	0.606	0.939	0.712	0.846	0.708
	LIME	4.3%	1.1%	1.8%	0.7%	0.845	0.631	0.779	0.614	0.940	0.702	0.820	0.695
Liu et al. [26]	None	5.8%	2.1%	3.1%	1.0%	0.826	0.649	0.722	0.652	0.909	0.725	0.812	0.719
Ours w/ loss traj.	SmoothGrad	16.3%	4.1%	7.3%	2.1%	0.876	0.652	0.856	0.619	0.961	0.750	0.915	0.708
	VarGrad	16.4%	3.8%	7.4%	1.5%	0.885	0.656	0.855	0.617	0.957	0.745	0.917	0.704
	IG	16.0%	3.8%	7.3%	1.7%	0.878	0.656	0.832	0.611	0.958	0.757	0.874	0.701
	GradCAM	15.8%	3.9%	8.4%	1.9%	0.887	0.656	0.854	0.616	0.962	0.751	0.914	0.700
	GradCAM++	16.5%	3.9%	7.3%	1.9%	0.897	0.632	0.854	0.616	0.962	0.750	0.914	0.700
	SHAP	15.7%	3.9%	7.2%	1.3%	0.861	0.652	0.803	0.624	0.961	0.755	0.864	0.708
	LIME	16.3%	4.0%	7.8%	1.3%	0.852	0.644	0.796	0.622	0.959	0.751	0.862	0.703

TABLE 3: Attack performance of our attack when combined with LiRA on CIFAR-10 datasets.

Model Num.	Method	TPR at 0.1% FPR	Balanced Acc.	AUC
16	LiRA [32]	1.10%	0.607	0.665
	Ours w/ LiRA	2.60%	0.611	0.686
256	LiRA [32]	7.20%	0.628	0.709
	Ours w/ LiRA	8.30%	0.632	0.715



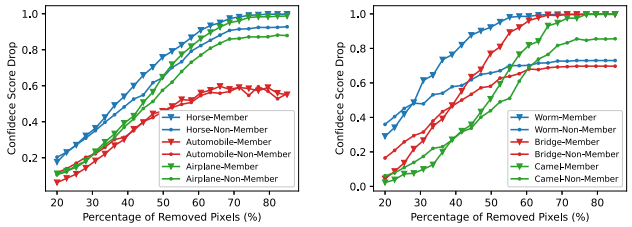
(a) Confidence Score (b) Confidence Score Drop

Figure 4: The perturbation trajectory for specific member and non-member samples that have similar small (< 0.01) losses on the model trained on CIFAR-10 and CIFAR-100.

the differences in privacy leakage levels across various classes and explanators.

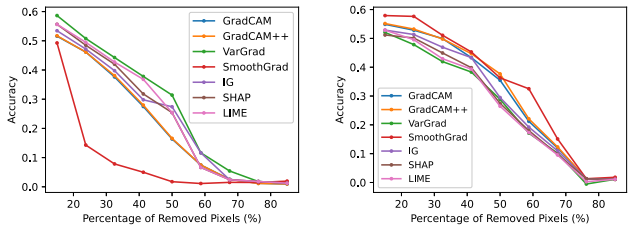
Root Causes of Privacy Leakage. The generalization gaps between training and testing data have been identified as major contributing factors to the success of membership inference attacks [22], [27]. These generalization gaps result in member samples having overall smaller losses than non-member samples, which are the primary attack features used by most previous work [22], [27], [31], [34]. However, these methods cannot differentiate between member and non-member samples with similar losses, leading to a high FPR. Our study shows that these generalization gaps also introduce differences in explanations, specifically, member samples are less robust against different levels of perturbations than non-member samples. These differences persist even when member and non-member samples exhibit similar losses. We employed GradCAM as the guiding explanation method for pixel removal, and Figure 4 shows the average confidence score and confidence score drop trajectory of member samples and non-member samples under MoRF settings when the loss is less than 0.01. In such cases, the loss-based methods cannot determine membership status; however, the perturbation trajectory still exhibits a clear difference between members and non-members. A possible explanation is that even in well-fitted samples with small losses, the features used for classification are not entirely identical. Member samples tend to rely more on key semantic information than non-member samples to derive a confident prediction. As a result, under different levels of perturbation, member samples may be less robust due to the removal of such key semantic information. Therefore, attribution maps can serve as complementary features alongside losses, significantly reducing the FPR of attacks.

Privacy Leakage across Different Classes. The extent of privacy leakage can vary across different classes of samples. Figure 5 displays the perturbation trajectories for three distinct classes in the CIFAR-10 and CIFAR-100 datasets under the MoRF setting. It is clear that each category presents a unique perturbation trajectory, resulting in varying degrees of difficulty in differentiating members and non-members. A possible explanation is that the model focuses on distinct attributes of images to infer specific classes. Furthermore, member samples and non-member samples differ in their utilization of such attributes for classification purposes, leading to varying levels of performance under perturbations.



(a) CIFAR-10 (b) CIFAR100

Figure 5: Comparison of perturbation trajectory between different classes in CIFAR10 and CIFAR-100.



(a) MoRF (b) LeRF

Figure 6: Comparison of explanation quality in MoRF and LeRF settings between different explanators.

Given this variation across classes, we incorporate the one-hot encoding of classes into our attack features to further enhance attack performance.

Privacy Leakage across Different Explanators. As illustrated in Figure 3 and Table 2, the degree of privacy leakage varies between different explanators. To investigate the possible reasons, we use the ROAD evaluation metrics [65] to assess the explanation quality in MoRF and LeRF settings among different explanators. As shown in Figure 6, different explanators exhibit distinct trends of change concerning different percentages of removed pixels. Specifically, SmoothGrad performs the best, demonstrating lower accuracy in MoRF and higher accuracy in LeRF. Grad-CAM++ and Grad-CAM are next in line, delivering fairly similar performances. Subsequently, IG, SHAP, and LIME are observed, with VarGrad lagging behind. Grad-CAM++ and Grad-CAM follow, achieving quite similar performances. Then IG, SHAP, and LIME follow, while VarGrad trails behind. An intriguing observation from Figure 3 and Table 2 is that our attack performs better with SmoothGrad, followed in order by Grad-CAM++, Grad-CAM, IG, SHAP, LIME, and VarGrad. This implies that *explanators with greater accuracy could potentially pose a higher risk of privacy leakage*, which leads to a privacy-utility trade-off consideration. The reason is that the more accurate explanators can offer more precise attributions to identify their focused features, thereby leading to wider generalization gaps between member and non-member samples.

Root Causes of Attack Effectiveness. Our investigation reveals significant privacy leakage. However, such leakage could stem from either the model explanations or our advanced attack strategy. To isolate the impact of explanations on privacy leakage, we designed two experiments. In the first experiment, we deviate from using precise explanations derived directly from the target model. Instead, we employ explanations from shadow models to conduct membership inference on the target model. This approach was evaluated on the CIFAR-100 dataset. The results given in Table 4 indicate a noticeable performance gap, which illustrates the extent of privacy leakage attributable to precise explanations as opposed to non-precise counterparts. For the second experiment, we followed our proposed attack method but excluded the use of explanations. Here, we implemented a pixel perturbation technique leveraging the super-pixel approach in LIME [24], where we grouped pixels into distinct segments based on similarities in color, intensity, or texture. Specifically, a quick shift algorithm [78] was employed for image segmentation. A selected percentage of these segments was then perturbed in line with our perturbation strategy, followed by the same membership inference process used in our method. The results given in Table 5 reveal a persistent performance gap when compared to attacks that utilize explanations, thereby revealing privacy leakage attributed to explanations. In addition, our explanation-free attack strategy demonstrated a notable improvement over baselines. This improvement could possibly stem from multiple queries in the inference process [79]. As a result, Table 2 shows the combined effect of the privacy leakage caused by explanations and an improved attack strategy.

TABLE 4: Attack performance of our method using explanations from shadow model on CIFAR-100 dataset.

Explanation Method	TPR at 0.1% FPR	Balanced Accuracy	AUC
SmoothGrad	3.7%	0.781	0.915
VarGrad	2.9%	0.708	0.926
IG	3.3%	0.732	0.917
GradCAM	2.8%	0.742	0.931
GradCAM++	3.2%	0.758	0.929
SHAP	2.2%	0.805	0.931
LIME	2.2%	0.782	0.919

TABLE 5: Attack performance of our method on different datasets without using explanations.

Dataset	TPR at 0.1% FPR	Balanced Accuracy	AUC
CIFAR100	3.9%	0.791	0.935
CIFAR10	0.6%	0.628	0.683
CINIC10	0.5%	0.773	0.804
GTSRB	0.6%	0.530	0.600

6.4. Ablation Study

Effects of Varied Attack Components. In this part, we investigate the effects of various components of our design on the overall attack performance by removing these parts individually, including the perturbation strategies, TV term, noisy linear imputation, and critical trajectory selection. More explicitly, we substitute MoRF and LeRF with a

TABLE 6: Ablation study of different attack components.

Variants	TPR at 0.1% FPR	Balanced Accuracy	AUC
w/o Perturb. Strategy	3.4%	0.773	0.918
w/o TV Term	4.0%	0.782	0.924
w/o Noisy Linear Imp.	3.1%	0.760	0.904
w/o Critical Traj. Sel.	2.9%	0.780	0.901
Full	5.0%	0.783	0.938

random selection (w/o Perturb. Strategy); we remove TV terms from equation 7 (w/o TV Term); we exchange the noisy linear imputation with imputation by the mean value per channel of the image (w/o Noisy Linear Imp.); and we employ a random trajectory selection (w/o Critical Traj. Sel.). The results are given in Table 6, serving to highlight the significance of each design choice in our method.

Using Solely Explanations. In this part, we focus on assessing the effectiveness of our attack by using solely explanations. Specifically, we employ confidence trajectories and the one-hot encoding of classes as the inputs for membership inference. The evaluation is conducted on the CIFAR-100 dataset. The results are given in Table 7, which demonstrate that our attack relying solely on explanations can still achieve notable performance.

Disjoint Datasets. In our previous experiments, we trained both the target models and shadow models from a common dataset, *i.e.*, we assume the adversary has the auxiliary dataset from the same distribution as the training dataset of the target model. In a more practical attack setting, the adversary likely has access to a dataset that is disjoint from the actual training data. We now show that this more practical setting has only a minor influence on the attack performance. For this experiment, we use the CINIC-10 dataset [58], which combines CIFAR-10 [57] with an additional 210k images from ImageNet that correspond to classes contained in CIFAR-10. Specifically, we consider two settings: in the first setting, the target model and shadow model are all trained on CIFAR-10 (*i.e.*, same distribution); while in the second setting, the target model is trained on the CIFAR-10 portion, and the adversary trained the shadow models on the ImageNet portion of CINIC-10. There is thus a distribution shift between the target model's dataset and the adversary's dataset. We use Grad-CAM to generate attribution maps in this experiment. As shown in Figure 7 and Table 8, our attacks' performance is not affected by the distribution shift of the datasets. Compared with the same distribution cases, there is only a 0.2% drop in TPR, and the balanced accuracy is even higher than that in the same distribution cases. Additionally, our attack performance is much higher than the baselines in this setting.

Different Model Architectures. After exploring the impact of the dataset distribution shift, we now relax another assumption of the adversary on the knowledge of the target model architecture. For this experiment, we vary the architectures of the shadow models with ResNet-18, VGG-16, ResNet-56, and DenseNet-161, while keeping the architectures of the target model as ResNet-18. As shown in Figure 8, our attack performs best when the target model and

TABLE 7: Attack performance using solely explanations.

Explanation Method	TPR at 0.1% FPR	Balanced Accuracy	AUC
SmoothGrad	2.3%	0.791	0.885
VarGrad	3.1%	0.823	0.889
IG	2.6%	0.709	0.825
GradCAM	3.2%	0.798	0.894
GradCAM++	2.8%	0.683	0.865
SHAP	2.7%	0.717	0.812
LIME	1.6%	0.768	0.842

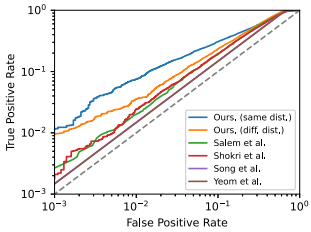


Figure 7: The ROC curve of different methods when two datasets are sampled from different distributions.

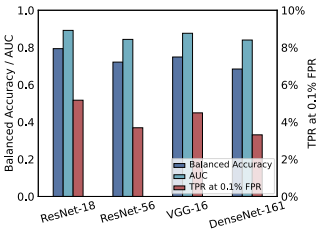


Figure 8: The impact of the architecture differences between the target model and shadow models.

shadow model share the same architecture. In addition, using models that have a similar number of layers (*e.g.*, ResNet-18, and VGG-16) will lead to a more similar performance. The possible reason is that the networks will similar depth will use similar features for classification, thus the generated attribution maps will be more similar in this scenario. When the architectures and number of layers are totally different, the attack performance will decrease, but the performance is still better other than the baselines using the same architecture for the shadow model and target model as shown in Table 2. To mitigate the performance loss, the adversary could try different shadow network architectures to find one that has a superior performance. These experiments have further demonstrated the severe privacy leakage caused by explanations, even when the adversary has limited knowledge of the target model architecture and training datasets.

Without Shadow Models. We explored how various model architectures influence attack efficacy with the above experiments. A notable challenge arises when the target models differ markedly from the shadow models, potentially diminishing the effectiveness of approaches based on shadow models. To address this, we adopt an alternative approach that eliminates the need for shadow models. Instead, we assume that an adversary has access to a portion of the training data for the target model, which could be implemented through data poisoning within the target model's training dataset [80]. In our experimental setup, the adversary holds non-members from the shadow model and a portion of members from the target model. This combined dataset is then used to train the attack model. The efficacy of this attack was assessed using the remaining member and non-member data of the target model. We varied the poisoning rate and evaluated the attack's performance on CIFAR-100,

TABLE 8: Comparison with different methods when shadow and target datasets are sampled from different distributions.

Method	TPR at 0.1% FPR	Balanced Accuracy	AUC
Yeom et al.	0.1%	0.606	0.656
Song et al.	0.1%	0.609	0.659
Salem et al.	0.3%	0.598	0.695
Shokri et al.	0.2%	0.592	0.686
Ours (Diff. Dist.)	1.0%	0.649	0.722
Ours (Same Dist.)	1.2%	0.647	0.722

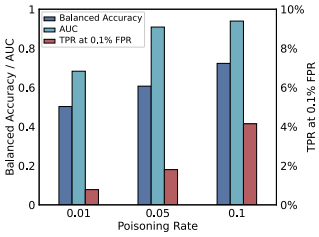


Figure 9: The impact of different poisoning rates on attack performance.

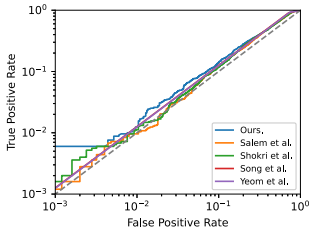


Figure 10: Comparison with baseline methods with small generalization gaps.

employing GradCAM as the explanation. The results given in 9 indicate that our attacks retained significant efficacy, particularly at higher poisoning rates.

Small Generalization Gaps. The effectiveness of membership inference attacks is closely tied to the generalization gaps of models. To this end, we investigate how our attack performs under small generalization gaps. To create such a scenario, we enriched the training dataset for the GTSRB and subsequently retrained our models. As a result, the trained target model exhibited a training accuracy of 99.87% while maintaining a testing accuracy of 90.88%. We then conducted membership inference attacks on this well-generalized model, utilizing GradCAM as the explanation. The results are shown in Figure 10. Our findings reveal a clear improvement of our approach over baseline methods, particularly within a low FPR range.

7. Possible Defenses

Considering the serious privacy concerns associated with explainable ML, we examine potential defenses and evaluate them empirically in this section. Although there is a broad spectrum of possible defenses [81], [82], [83], [84], [85], [86], we consider from two angles: the model level and the output level. At the model level, we use differential privacy [82] to build an inherently private model. At the output level, we use MemGuard [83] to strategically disrupt the model's output, thereby restricting the information accessible to attackers.

Differential Privacy. Differential privacy (DP) is a widely used mechanism to prevent classical membership inference attacks [80]. Essentially, this approach imposes a constraint on the ability to distinguish between two adjacent datasets that differ solely by the presence or absence of one data sample. As a result, DP plays a crucial role in safeguarding

TABLE 9: Attack performance of our attack against DP-SGD under different privacy budgets.

Privacy Budget (ε)	Top-1 Accuracy Drop	Top-5 Accuracy Drop	TPR at 0.1% FPR	Balanced Accuracy
200000	0.113	0.164	0.2%	0.544
1000	0.288	0.278	0.1%	0.511
100	0.347	0.375	0.1%	0.506

against membership inference. We adopt the Differentially-Private Stochastic Gradient Descent (DP-SGD) [82], the most representative DP mechanism for protecting the models. We use the recent Fast Differential Privacy library [87] to implement the DP framework. Specifically, we configure the gradient clipping function to 'automatic' and select 'MixOpt' as the clipping model. This clipping process is uniformly applied across all layers of our model. We use ResNet-18 trained on CIFAR-100 for this experiment and use Grad-CAM as the explanation method. As shown in Table 9, although DP-SGD could decrease our attack performance, it will decrease the classification accuracy significantly even when ϵ is large. We also evaluate how DP-SGD will affect the explanation quality. Specifically, we use ROAD evaluation metrics to evaluate the quality of attribution maps at $\epsilon = 200000$. Additionally, we introduce a baseline called Sobel edge filter following [62], which generates the attribution maps by assigning a high score to areas of the image with a high gradient. This process does not depend at all on the original model parameters. Therefore, this baseline represents a lower bound in performance that all explanation methods are expected to outperform. The attribution quality evaluation is given in Figure 11 in Appendix. We can observe that the accuracy in the Sobel edge filter decreases more sharply in MoRF settings while remaining high in the LeRF setting. This means the generated attribution maps are not even more informative than methods that even do not take into account the model parameters. Therefore, it is hard to balance the trade-off between defense capability and performance utility.

MemGuard. In contrast to the defense strategies that alter the training process, Jia et al. [83] introduced MemGuard. This technique strategically introduces perturbations into the confidence scores of the target models for each input. Specifically, the methodology is designed to convert the perturbed confidence scores into adversarial examples for the attack models, which will confound the adversary by generating random membership inference results based on the perturbed score vector. To implement MemGuard, we adhered to the default parameters in the original paper and set a confidence score distortion budget to 1.0 to guarantee the robust defenses. For this experiment, we employed a ResNet-18 trained on the CIFAR-100 dataset and utilized SmoothGrad, Grad-CAM, LIME, and SHAP as the explanation methods. The results, as depicted in Table 10, reveal that our attacks are still effective, even in the presence of such stringent defenses. The underlying reason is that MemGuard solely perturbs the target model's output through the addition of noise, leaving the attribution maps unprotected.

TABLE 10: Attack performance of our attack against MemGuard for different explanation methods.

Explanation Method	TPR at 0.1% FPR	Balanced Accuracy	AUC
SmoothGrad	2.6%	0.778	0.885
GradCAM	2.4%	0.775	0.880
SHAP	2.3%	0.761	0.877
LIME	2.3%	0.769	0.872

Consequently, it fails to shield against the employed attacks.

8. Related Work

Explainable Machine Learning. Explainable ML can be easily adapted to various network architectures without altering the original models, making it widely used for explaining complex black-box neural networks [9], [40]. Among various explanation methods, attribution-based methods that have become the most studied in recent years [4], [6], [41], [42], [43], [46], as they can generate a high-fidelity pixel-wise explanation of the model decision-making process. Explainable ML has shed light on the inner workings of models by making their decision-making processes more understandable to humans. As a result, it is believed to provide better insight into the correctness and robustness of these models [88]. Existing research has applied explainable ML to numerous security and privacy-sensitive domains, such as medical diagnosis [12], [13], privacy protection in voice assistants [89], and anomaly detection [10].

Membership Inference Attack. Membership inference attack has gained increasing attention from academia [22], [26], [27], [31], [32], [55], [90]. It has been utilized as a basic auditing tool to quantify the privacy leakage of deep learning models in different scenarios [91]. Liu et al. [92] focus on the membership inferences on pre-trained encoders in contrastive learning. Chen et al. [54] conduct membership inference against GANs with different levels of knowledge. Chen et al. [33] investigate the membership inference in machine unlearning, where they use extracted features from original models and unlearned models to infer the membership status. Yuan et al. [28] investigate the membership inference against neural network pruning, where they use prediction sensitivity and confidence to distinguish the membership status. Li et al. [29] conduct membership inference against multi-exit neural networks. They use the additional exit information to help with inferring to achieve better attack performance. However, membership inference attacks on explainable machine learning are less investigated.

Security of Explainable ML. Machine learning models have been recognized as susceptible to a variety of security risks [93], [94], [95]. Recently, investigating the security risks of explainable ML has been an emerging research field in machine learning security [15], [16], [17], [18], [96]. Dombrowski et al. [16] show that attribution maps can be easily manipulated by applying imperceptible perturbations to the images without affecting the prediction output. Instead of perturbing the input data, Heo et al. [17] found out that the explanation methods can be fooled on the entire validation data set by fine-tuning a pre-trained network. Zhang et al. [18] perform a systematic study of the security of deep learning explanators, and they show adversarial examples can fool both the target DNNs and their coupled explanators, simultaneously exposing the adversarial vulnerabilities of existing explanation models.

Privacy of Explainable ML. While the security risks associated with explainable ML have received significant attention, the inherent privacy risks in explainable ML remain insufficiently examined, with only a handful of studies exploring this critical issue. Shokri et al. [20] revealed vulnerabilities in backpropagation-based explanations, demonstrating that the variance in these explanations can reveal membership status. Pawelczyk et al. [97] introduced a membership inference attack specifically designed for counterfactual (CF) explanations, utilizing the distances between data instances and their CF counterparts. In addition, Luo et al. [21] focused on feature inference attacks by reconstructing private inputs based on their Shapley value explanations, while Duddu et al. [98] developed an attribute inference attack to deduce sensitive attributes, such as race and sex, using model explanations. Furthermore, Zhao et al. [99] proposed an attack leveraging explanations as additional features to enhance model inversion attacks. Beyond these data privacy concerns, there have been efforts using explanations to facilitate model extraction attacks due to their ability of revealing the model's decision boundary [100]. In response to these privacy concerns, there have also been studies in countermeasures using differential privacy [101], [102]. However, despite these developments, a systematic investigation into the privacy risks of attribution-based explanations through the lens of membership inference, remains underexplored.

9. Discussion

The Impacts of Our Attack. Transparent and accurate explanations are crucial in building trust in ML systems, especially in critical domains such as computer security [103]. Our study reveals the privacy implications of incorporating explanations in ML systems, thus empowering system maintainers and developers to consider these risks when making deployment decisions for these techniques. Moreover, regulators and policymakers can be aware of these risks when formulating rules and guidelines for ML systems. Our findings also serve as a guide for security researchers to further investigate the privacy risks associated with explainable ML and to develop effective defenses. When all participants in the ML ecosystem consider these privacy risks, we can move towards a more secure and reliable ML landscape.

Advanced Defense Mechanisms. Our investigation of differential privacy as a defense mechanism shows that while it effectively thwarts privacy breaches, it significantly impacts classification accuracy and explanation quality. On the other hand, MemGuard has proven to be inadequate in defending

against privacy breaches. Hence it is crucial to devise more advanced defense mechanisms to counter such evolving threats, which we propose as a future research direction.

10. Conclusion

In this paper, we present a systematic study on the privacy leakage of explainable machine learning through the lens of membership inference. Building on the observation that robustness exhibits observable differences among member samples and non-member samples, we propose a novel method to extract the membership features from changes in model confidence under different levels of perturbations guided by the attribution maps. We evaluated our approach with seven popular explanators across various benchmarks and model architectures. The evaluation results show that our attacks consistently outperform prior works, highlighting significant privacy leakage in current explainable ML methods. We have further demonstrated that this leakage issue persists even when the attacker lacks knowledge of training datasets or target model architectures. Finally, we empirically evaluate the existing model and output-based defense, finding that they are insufficient in mitigating our attacks. We hope this study will raise awareness of the potential privacy risks of using explainable machine learning and help improve privacy in practical implementation.

Acknowledgment

We thank the reviewers for their valuable feedback. This work was partially supported by the NSF (CNS-1916926, CNS-2229427, CNS-2238635), and ARO (W911NF2010141), and Washington University.

References

[1] "Chatgpt," https://openai.com/blog/chatgpt, 2022.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[3] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[5] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[6] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[7] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[8] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[9] A. S. Jacobs, R. Beltiukov, W. Willinger, R. A. Ferreira, A. Gupta, and L. Z. Granville, "Ai/ml for network security: The emperor has no clothes," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 1537–1551.

[10] D. Han, Z. Wang, W. Chen, Y. Zhong, S. Wang, H. Zhang, J. Yang, X. Shi, and X. Yin, "Deepaid: interpreting and improving deep learning-based anomaly detection in security applications," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 3197–3217.

[11] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "Lemna: Explaining deep learning based security applications," in *proceedings of the ACM SIGSAC conference on computer and communications security*, 2018.

[12] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature biomedical engineering*, vol. 2, no. 10, pp. 749–760, 2018.

[13] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[14] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[15] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[16] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," *Advances in neural information processing systems*, vol. 32, 2019.

[17] J. Heo, S. Joo, and T. Moon, "Fooling neural network interpretations via adversarial model manipulation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang, "Interpretable deep learning under fire," in *29th {USENIX} security symposium ({USENIX} security 20)*, 2020.

[19] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[20] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 231–241.

[21] X. Luo, Y. Jiang, and X. Xiao, "Feature inference attack on shapley values," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2233–2247.

[22] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, "Membership inference attacks by exploiting loss trajectory," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2085–2098.

[27] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[28] X. Yuan and L. Zhang, "Membership inference attacks and defenses in neural network pruning," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4561–4578.

[29] Z. Li, Y. Liu, X. He, N. Yu, M. Backes, and Y. Zhang, "Auditing membership leakages of multi-exit networks," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[30] L. Song, R. Shokri, and P. Mittal, "Privacy risks of securing machine learning models against adversarial examples," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 241–257.

[31] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models." in *USENIX Security Symposium*, vol. 1, no. 2, 2021, p. 4.

[32] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[33] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 896–911.

[34] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.

[35] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[36] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794*, 2017.

[37] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[38] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[39] R. Marcinkevičs and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," *arXiv preprint arXiv:2012.01805*, 2020.

[40] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[41] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[42] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[43] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3429–3437.

[44] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[45] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[46] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[47] J. Seo, J. Choe, J. Koo, S. Jeon, B. Kim, and T. Jeon, "Noise-adding methods of saliency map as series of higher order partial derivative," *arXiv preprint arXiv:1806.03000*, 2018.

[48] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[49] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," *arXiv preprint arXiv:2111.09679*, 2021.

[50] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial intelligence in medicine*, vol. 94, pp. 42–53, 2019.

[51] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable ai in fintech risk management," *Frontiers in Artificial Intelligence*, vol. 3, p. 26, 2020.

[52] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Explainable machine learning for fake news detection," in *Proceedings of the 10th ACM conference on web science*, 2019, pp. 17–26.

[53] M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, and Y. Zhang, "Membership inference attacks against recommender systems," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 864–879.

[54] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 2020, pp. 343–362.

[55] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 880–895.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] "Cifar dataset," https://www.cs.toronto.edu/~kriz/cifar.html, 2012.

[58] L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey, "Cinic-10 is not imagenet or cifar-10," *arXiv preprint arXiv:1810.03505*, 2018.

[59] "Gtsrb dataset," https://benchmark.ini.rub.de/?section=gtsrb, 2011.

[60] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[62] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[63] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[64] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 6021–6029.

[65] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for attribution methods," *arXiv preprint arXiv:2202.00449*, 2022.

[66] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," *arXiv preprint arXiv:1706.07206*, 2017.

[67] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.

[68] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[69] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[70] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[71] B. L. Welch, "The generalization of 'student's' problem when several different population varlances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.

[72] F. E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics bulletin*, vol. 2, no. 6, pp. 110–114, 1946.

[73] X. He and Y. Zhang, "Quantifying and mitigating privacy risks of contrastive learning," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 845–863.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[75] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[76] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems*, vol. 4, 1991.

[77] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[78] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*. Springer, 2008, pp. 705–718.

[79] Y. Wen, A. Bansal, H. Kazemi, E. Borgnia, M. Goldblum, J. Geiping, and T. Goldstein, "Canary in a coalmine: Better membership inference with ensembled adversarial queries," *arXiv preprint arXiv:2210.10750*, 2022.

[80] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, "{ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4525–4542.

[81] S. Kundu, Y. Zhang, D. Chen, and P. A. Beerel, "Making models shallow again: Jointly learning to reduce non-linearity and depth for latency-efficient private inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4684–4688.

[82] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[83] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 259–274.

[84] Y. Xiao, N. Zhang, J. Li, W. Lou, and Y. T. Hou, "Privacyguard: Enforcing private data usage control with blockchain and attested off-chain contract execution," in *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part II 25*. Springer, 2020, pp. 610–629.

[85] J. Wang, Y. Wang, and N. Zhang, "Secure and timely gpu execution in cyber-physical systems," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 2591–2605.

[86] J. Wang, A. Li, H. Li, C. Lu, and N. Zhang, "Rt-tee: Real-time system availability for cyber-physical systems using arm trustzone," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 352–369.

[87] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Differentially private optimization on large model at small cost," in *International Conference on Machine Learning*. PMLR, 2023, pp. 3192–3218.

[88] A. Nadeem, D. Vos, C. Cao, L. Pajola, S. Dieck, R. Baumgartner, and S. Verwer, "Sok: Explainable machine learning for computer security applications," *arXiv preprint arXiv:2208.10605*, 2022.

[89] Y. Chen, Y. Bai, R. Mitev, K. Wang, A.-R. Sadeghi, and W. Xu, "Fakewake: Understanding and mitigating fake wake-up words of voice assistants," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 1861–1883.

[90] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[91] Z. Zhang and R. Hu, "Byzantine-robust federated learning with variance reduction and differential privacy," in *2023 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2023, pp. 1–9.

[92] H. Liu, J. Jia, W. Qu, and N. Z. Gong, "Encodermi: Membership inference against pre-trained encoders in contrastive learning," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2081–2095.

[93] H. Liu, Z. Yu, M. Zha, X. Wang, W. Yeoh, Y. Vorobeychik, and N. Zhang, "When evil calls: Targeted adversarial voice over ip network," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2009–2023.

[94] H. Liu, Y. Wu, Z. Yu, Y. Vorobeychik, and N. Zhang, "Slowlidar: Increasing the latency of lidar-based detection using adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5146–5155.

[95] H. Liu, Y. Wu, S. Zhai, B. Yuan, and N. Zhang, "Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 585–20 594.

[96] Z. Chen, F. Silvestri, J. Wang, Y. Zhang, and G. Tolomei, "The dark side of explanations: Poisoning recommender systems with counterfactual examples," *arXiv preprint arXiv:2305.00574*, 2023.

[97] M. Pawelczyk, H. Lakkaraju, and S. Neel, "On the privacy risks of algorithmic recourse," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 9680–9696.

[98] V. Duddu and A. Boutet, "Inferring sensitive attributes from model explanations," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 416–425.

[99] X. Zhao, W. Zhang, X. Xiao, and B. Lim, "Exploiting explanations for model inversion attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 682–692.

[100] Y. Wang, H. Qian, and C. Miao, "Dualcf: Efficient model extraction attack from counterfactual explanations," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1318–1329.

[101] N. Patel, R. Shokri, and Y. Zick, "Model explanations with differential privacy," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1895–1904.

[102] F. Yang, Q. Feng, K. Zhou, J. Chen, and X. Hu, "Differentially private counterfactuals via functional mechanism," *arXiv preprint arXiv:2208.02878*, 2022.

[103] Z. Yu, S. Zhai, and N. Zhang, "Antifake: Using adversarial audio to prevent unauthorized speech synthesis," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 460–474.

Appendix A. Preliminary Study

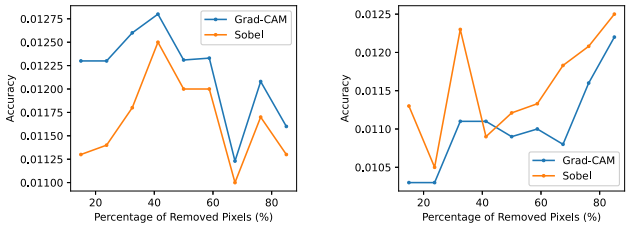
Our preliminary study on membership inference against explainable ML directly employs attribution maps as features to identify membership. This can be considered a standard image classification task comprising one positive class (*i.e.*, members) and one negative class (*i.e.*, non-members). We use ResNet-18 as the target model and shadow model architecture, and we divide the datasets as described in Section 6.1 to train the shadow models and target models. ResNet-50 serves as the attack model to classify membership status. Grad-CAM is utilized to generate attribution maps. The attack results can be found in Table 11.

TABLE 11: Attack performance by directly classifying membership status from attribution maps.

Dataset	TPR at 0.1% FPR	Balanced Accuracy	AUC
CIFAR-10	0.0%	0.499	0.498
CIFAR-100	0.1%	0.502	0.498
CINIC-10	0.1%	0.503	0.499
GTSRB	0.1%	0.500	0.501

Appendix B. Experimental Settings

Dataset Splits. To evaluate attack performance in various settings, we follow [26] to split the datasets. The detailed data splits for different datasets are provided in Table 12. $\mathcal{D}_{target}^{train}$ is used to train the target models, and $\mathcal{D}_{shadow}^{train}$ is used to train the shadow models; they are considered member datasets, while $\mathcal{D}_{target}^{test}$ and $\mathcal{D}_{shadow}^{test}$ are considered non-member datasets. $\mathcal{D}_{distill}$ is only used for the second setting, where the adversary has access to additional large datasets and employs them for model distillation.



(a) MoRF (b) LeRF

Figure 11: Comparison of explanation quality in MoRF and LeRF settings between Grad-CAM and Sobel Edge Filter when the model is trained on DP-SGD.

TABLE 12: Data splits on different datasets.

Dataset	$\mathcal{D}_{target}^{train}$	$\mathcal{D}_{target}^{test}$	$\mathcal{D}_{shadow}^{train}$	$\mathcal{D}_{shadow}^{test}$	$\mathcal{D}_{distill}$
CIFAR-10	10000	10000	10000	10000	20000
CIFAR-100	10000	10000	10000	10000	20000
CINIC-10	10000	10000	10000	10000	220000
GTSRB	1500	1500	1500	1500	45837

Training Configurations. We train each model for 100 epochs with a learning rate of 0.1. We also reduce the learning rate of the optimizer in a cosine annealing schedule to ensure better model convergence. Standard data augmentations and weight decays with rate of 0.0001 are used to improve the generalization of the models.

Model Accuracy. Table 13 shows the performance of the target ResNet-18 models trained on different datasets.

TABLE 13: Training and testing accuracy for the target model on different datasets.

Dataset	Top1 Train Acc	Top1 Test Acc	Top5 Train Acc	Top5 Test Acc
CIFAR-10	0.998	0.771	1.000	0.982
CIFAR-100	1.000	0.454	1.000	0.739
CINIC-10	0.998	0.590	1.000	0.938
GTSRB	0.997	0.745	1.000	0.947

TABLE 14: Performance of baseline methods utilizing augmentations during attacks.

Method	Dataset	TPR at 0.1% FPR	Balanced Accuracy	AUC
Salem et al.	CIFAR100	0.8%	0.851	0.904
	CIFAR10	0.2%	0.595	0.623
	CINIC10	0.3%	0.705	0.760
	GTSRB	0.4%	0.608	0.666
Shokri et al.	CIFAR100	1.1%	0.780	0.853
	CIFAR10	0.3%	0.604	0.638
	CINIC10	0.3%	0.708	0.764
	GTSRB	0.3%	0.611	0.681

Appendix C.
Meta-Review

The following meta-review was prepared by the program committee for the 2024 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

C.1. Summary

The paper presents an empirical study of the privacy risks of explainable machine learning. The study uses a new membership inference attack strategy that uses prediction trajectories derived from explanations.

C.2. Scientific Contributions

- Addresses a Long-Known Issue
- Provides a Valuable Step Forward in an Established Field.

C.3. Reasons for Acceptance

1) The paper addresses a long-known issue. The need to study privacy risks of explanations is evident in the existing literature. The paper addresses this through extensive experiments.
2) The paper provides a valuable step forward in an established field. Prior studies have investigated privacy risks of explanations using membership inference attacks. The paper proposes and evaluates a stronger attack strategy to quantify the privacy leakage of explanations.