Explanations Leak: Membership Inference with Differential Privacy and Active Learning Defense

Fatima Ezzeddine[†,*], Osama Zammar [‡], Silvia Giordano[†] and Omran Ayoub[†]
[†] *University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland*
[*] *Università della Svizzera italiana, Lugano, Switzerland*
[‡] *Lebanese University, Faculty of Science, Beirut, Lebanon*

***Abstract*—Counterfactual explanations (CFs) are increasingly integrated into Machine Learning as a Service (MLaaS) systems to improve transparency. However, ML models deployed via Application Programming Interfaces (APIs) are already susceptible to privacy attacks, such as membership inference and model extraction, and the implications of providing explanations on this threat landscape remains largely unexplored. In this work, we investigate how CFs inadvertently expand the attack surface of MLaaS by amplifying the effectiveness of membership inference attacks (MIAs). First, we systematically analyze how exposing CFs through query-based APIs enables more effective shadow-based MIAs. Second, we propose a defense framework that integrates Differential Privacy (DP) with Active Learning (AL) to jointly reduce memorization and limit effective training data exposure. Finally, we conduct an extensive empirical evaluation to characterize the three-way trade-off between privacy leakage, predictive performance, and explanation quality. Our findings highlight the need to carefully balance transparency, utility, and privacy in the responsible deployment of explainable MLaaS.**

***Index Terms*—Counterfactual Explanations, Differential Privacy, Active Learning, Membership Inference Attack**

I. Introduction

Explainable AI (XAI) methods are increasingly employed in Machine Learning as a Service (MLaaS) systems with the aim of offering users and stakeholders more transparency into the decision-making process of deployed Machine learning (ML) models [1]–[3]. XAI frameworks provide insights into the factors underlying a model's prediction, allowing stakeholders, users, or regulators interpret automated decisions. For example, feature-importance methods attribute predictive influence to input variables, identifying which features drive a decision and how changes in their values affect model outputs [4]. Counterfactual explanations (CFs) demonstrate how minimal input changes can alter a model's decision in a hypothetical scenario, highlighting which factors drive predictions [5].

In standard MLaaS deployments, models are exposed via remote Application Programming Interfaces (APIs) that provide queriable prediction functionality. Through this query interface, users and stakeholders submit their data and receive the model's outputs. While such API access is essential for MLaaS, it also creates a broad attack surface that can be exploited to compromise privacy by leveraging the model's responses. Several privacy attacks operate in this setting, auch as Membership inference attacks (MIAs) that aim to determine whether a specific record was included in the model's training set, model extraction attacks that attempt to replicate the behavior of the target model, and property inference attacks that seek to infer sensitive aggregate properties of the training data beyond the model's intended disclosures [3]. Integrating explanations into MLaaS, where, in addition to predictions, models also return explanations, introduces new security and privacy concerns. Recent research has further demonstrated that providing explanations alongside predictions can amplify these risks [6]. This is because explanations, and particularly CFs, often reveal features, dependencies, and boundary regions on which the model relies, they provide adversaries with additional signals that can be systematically leveraged to infer sensitive information about the underlying training dataset, thereby increasing the effectiveness and precision of privacy attacks [2], [3], [6]–[8].

Among these threats, MIAs are particularly critical because they directly undermine the confidentiality of training data. By revealing whether a specific individual's record was used to train a model, MIAs create a tangible link between the deployed model and identifiable individuals. Such leakage significantly increases the risk that the model itself may be considered personal data under the General Data Protection Regulation (GDPR), since it can enable the re-identification of individuals [9]. This concern is especially acute when models are trained on sensitive datasets, where confirming membership may expose private attributes or sensitive participation. If explanations are available, attackers can further strengthen MIAs by combining standard model outputs (e.g., predicted probability distributions) with additional information disclosed through explanations. In particular, CFs can reveal how far an instance lies from the model's decision boundary and can be exploited by the intuition that training samples are more confidently classified and thus are often located farther from the decision boundary than non-members. Consequently, an adversary may infer membership when the distance between a given instance and its CF exceeds a pre-defined threshold [7].

In this context, it is essential to systematically analyze how CFs can be exploited to strengthen privacy attacks, such as MIAs, and to quantify the extent to which they enhance their effectiveness. At the same time, developing robust defense mechanisms is equally critical. Privacy-enhancing technologies naturally emerge as candidate solutions, and recent research has explored the application of Differential Privacy (DP), either to the underlying ML models or directly to the explanations [10], [11]. However, introducing privacy-

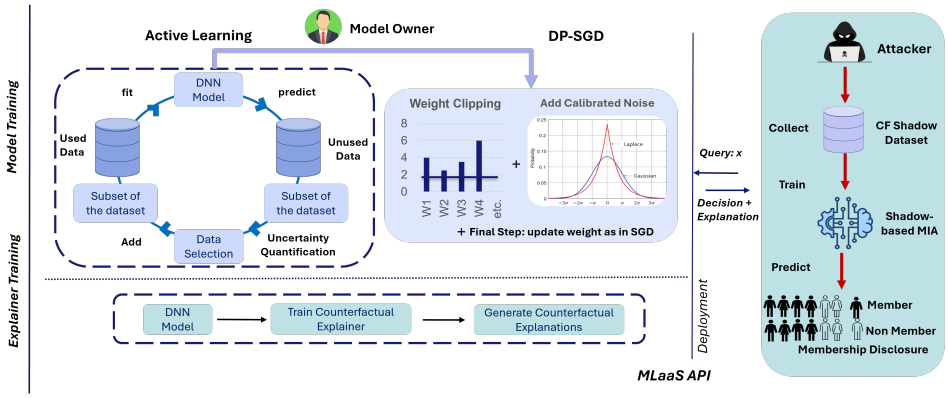

Fig. 1: Scenario illustrating model training procedures performed, alongside the attacker's strategy of performing MIA.

preserving mechanisms is not without cost. Applying DP may affect not only the model's predictive performance but also the usefulness of the generated explanations [12]. Therefore, it is necessary to carefully evaluate how privacy protection impacts explanation utility, alongside model predictive performance. This tension gives rise to a fundamental interplay between three competing objectives: (i) resilience against privacy attacks such as MIAs, (ii) predictive performance of the underlying ML model, and (iii) the quality of the provided explanations.

In this work, we address this three-way tension, with the broader goal of contributing to the responsible deployment of XAI in socially sensitive high-stakes domains such as finance and public services. It is imperative to ensure that efforts to enhance transparency do not expose individuals to privacy harms. To this end, we first investigate how CFs can be exploited to perform MIAs, in which an adversary has API query access to both predictions and CFs. In particular, we focus on *shadow-based MIAs*, a practical, realistic, and well-established attack methodology, compared with *threshold-based MIAs* [13], which rely on predefined thresholds of membership indicators such as loss or confidence. Specifically, in *shadow-based MIAs*, attackers approximate the target model's behavior using surrogate models trained on auxiliary data. This assumption reflects realistic attack capabilities in deployment. Second, we propose a novel defense mechanism that integrates DP with Active Learning (AL)[1] as a data-distillation strategy [14]–[18]. Our approach seeks to jointly combine DP and AL in a manner to strengthens resistance to MIAs while preserving both predictive performance and explanation quality (see Fig. 1). This design enables an evaluation of the trade-off between privacy preservation, measured by the attacker's ability to correctly infer membership, and explainability, quantified by CF quality.

To the best of our knowledge, this is among the first works to conduct a comprehensive study on shadow-based MIAs that leverage CFs and to propose a defense mechanism that combines DP and AL. Moreover, we systematically analyze the interplay between model predictive performance, attack success, and CF quality. The contributions can be summarized as:

- We systematically quantify the extent to which CFs can be leveraged as the adversary's knowledge source for shadow-based MIA.
- We introduce a novel defense framework that integrates DP with AL to limit the model memorization of training data and reduce training data exposure.
- We empirically characterize the three-way trade-off between privacy leakage, predictive performance, and CF quality, offering a unified analysis of privacy, utility, and explanation fidelity.

II. Related Work

Explanation-based MIA have emerged as powerful methods to perform threshold-based MIA [19]. For instance, [3] demonstrates that back-propagation-based explanations strengthen MIA by leveraging the intuition that higher variance in feature attributions indicates greater uncertainty in the logits, signaling that an instance is less likely to be a training member. [7] introduced an MIA recourse-based attack using CFs, exploiting the intuition that training points should be farther from decision boundaries than test points, predicting membership when the distance to a CF exceeds a threshold. Authors in [20] further extend this understanding by introducing robustness-based MIA, where perturbations guided by attribution maps reveal that members exhibit confidence drops when important features are altered.

Other efforts have focused on defending against MIA [21], [22]. In [21], the authors propose regularization-based defenses and attribute the success of MIAs to overfitting. The authors of [23] suggest overfitting-prevention techniques such as dropout and L2 regularization. Although helpful, these methods are not specifically designed for MIA defense and therefore lack robustness. Other methods have been designed specifically to defend against MIA. Authors in [22] introduced an adversarial regularization method that adds a gain of membership inference term to the loss function, effectively creating a max-min game where the model maximizes and minimizes this term during training. [24] proposed a regularizer between the softmax output distributions of the training and validation sets to prevent MIA. Another type of defense is adversarial

[1]Our intuition to use AL as a technique to further enhance privacy is discussed in more detail in Sec. III

example-based, which intentionally alters inputs or outputs, often through perturbations, to mislead an attacker. For instance, [25] proposed adding crafted noise to the model's outputs, thereby transforming them into adversarial examples that mislead the attacker's classifier and prevent MIA.

Several other works have explored the use of DP to defend against MIA. [26] proposed the DP-Stochastic Gradient Descent algorithm (DP-SGD), which builds on enhancing the original stochastic gradient descent by integrating DP into the backpropagation step. The work in [27] revealed that even DP-SGD could still be compromised by MIA methods. To mitigate the drawbacks of employing DP during model training or data generation, other works have explored incorporating DP at different stages. For instance, [28] modifies and normalizes confidence score vectors via a DP mechanism, thereby safeguarding privacy by obscuring membership information and reconstructed data, while [29] provided an in-depth analysis of privacy risks in neural network pruning and discovered susceptibility to MIAs. In the interplay between XAI and privacy, the demonstrated tension between them has prompted research into methods that balance explainability with privacy guarantees [30]. Studies reveal an inherent trade-off: applying privacy-preserving techniques such as DP to models and explainers often degrades the quality and utility of the explanations [12], [31]. In [30], it is demonstrated that significant visual degradation occurs in Deep Neural Network (DNNs) explanations when DP is applied [32]. Despite these challenges, privacy-preserving xAI approaches have emerged. Authors in [33] successfully integrate DP into Explainable Boosting Machines, achieving promising accuracy with formal privacy guarantees, with the limitation of being model-specific.

In contrast to prior work, this is the first work to investigate how CFs can be exploited within a shadow-based MIA framework, positioning CFs as a primary source of adversarial knowledge. Moreover, our work is the first to examine the combined use of DP and AL as a defense strategy against MIAs. While prior studies have primarily applied AL to reduce labeling costs or to generate synthetic datasets, we integrate AL with DP and assess how this combination strengthens privacy in MLaaS settings. The objective is to offer a comprehensive analysis of their interaction and its impact on model utility, membership leakage, and CF quality.

III. Reference Scenario and Methodology

A. Reference Scenario: Shadow-based MIA

Figure 1 illustrates our reference scenario. Consider a DNN model f_θ trained on a dataset $\mathcal{D}$: $\mathbf{X} \in \mathbb{R}^d \rightarrow \mathbf{Y} \in \mathbb{R}^t$. The model f_θ takes as input a numerical query $\mathbf{x} \in \mathbf{X}$ and predicts an output $\mathbf{y} \in \mathbf{Y}$. Let θ denote the weight matrix and $\mathbf{b}$ the bias vector of the DNN. The vector $\mathbf{y} = f_\theta(\mathbf{x})$ represents confidence scores (probabilities), indicating the model's probability toward each class. The model f_θ is deployed as an MLaaS, and the model owner trains a CF explainer E. The CF explainer E is trained on the dataset $\mathcal{D}$ to generate CF instances. The MLaaS returns $E(\mathbf{x}), f_\theta(\mathbf{x})$ when queried by x. An attacker aims to determine whether specific data points belong to $\mathcal{D}$ via MIA by exploiting the MLaaS API's output, as described in [23].

Specifically, the attacker collects a shadow dataset by repeatedly querying the API and distributes it into several datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_S\}$, each drawn from the same distribution $\mathcal{P}$ as $\mathcal{D}_{\text{train}}$. For each dataset $\mathcal{D}_s$, the adversary trains a set of *shadow models* $f_s(\cdot;\theta_s)$, aiming to approximate the behavior of the target model $f(\cdot;\theta)$. The adversary fully controls $\mathcal{D}_s$ and its membership status, so for every sample $x \in \mathcal{D}_s$, it is known that x is a *member*, while for $x \notin \mathcal{D}_s$ (but drawn from $\mathcal{P}$), it is a *non-member*. Each shadow model f_s returns a posterior vector $\boldsymbol{p}_s(x) = f_s(x;\theta_s) \in \mathbb{R}^K$, where K is the number of classes. Using the shadow models posteriors $\boldsymbol{p}_s(x)$ for both *member* and *non-member* samples, the adversary trains an *attack model* $g(\cdot;\phi)$ to infer membership:

$$g\big(\boldsymbol{p}_s(x);\phi\big) = \begin{cases} 1, & \text{if } x \in \mathcal{D}_s, \\ 0, & \text{otherwise.} \end{cases}$$

We evaluate the MIA across two configuration settings that differ in how the shadow dataset is constructed: i) No-CF setting and ii) CF-enabled setting. The collected dataset is then used to train the shadow models and indicates whether the MLaaS provider offers CF explanations. In both cases, the attacker must first assemble a shadow dataset that approximates the distribution of the original training data. Specifically:

a) No-CF setting: The attacker does not have access to CF explanations, we follow the standard assumption in the literature: the shadow dataset contains model outputs for inputs drawn from the same distribution as the training data.

$$\mathcal{D}_{\text{shadow}} = \{f_\theta(\mathbf{x}_i)\}_{i=1}^{M},$$

b) CF-enabled setting.: The attacker leverages CF explanations provided by the MLaaS system, queries the CF explainer E for a set of input instances, and receives both the model outputs and CF explanations, thereby forming a shadow dataset. Access to CFs provides the attacker with additional side information $E(x)$ that increases the mutual information between the model's outputs and the membership variable.

$$\mathcal{D}_{\text{shadow}} = \{E(\mathbf{x}_i),\ f_\theta(\mathbf{x}_i)\}_{i=1}^{M}.$$

Therefore, the shadow model trained on $\mathcal{D}_{\text{shadow}}$ resulting in attack model $g(p_s(x) \parallel \phi(E(x), x)$. Note that while baselines such as CF-distance threshold exist, we do not include them because their performance depends strongly on the chosen CF generator and on feature scaling/normalization, making fair reproduction and threshold calibration ambiguous.

B. Defense with Differential Privacy and Active Learning

To defend against MIA, we formulate two main objectives (i) to limit the model's tendency to memorize training data during training, and (ii) to reduce the amount and sensitivity of data exposed. Therefore, we propose a defense mechanism where, during the model training phase, the model owner leverages both AL and DP. Specifically, we employ AL, despite being traditionally used to reduce annotation costs by

selecting only the most informative samples, as a technique to train the DNN with the aim of reduce training data by constructing a smaller yet high-utility training data subset. Moreover, we employ DP, specifically, DP-SGD on the model level with rigorous privacy accounting to achieve a target ε-DP budget, to provide formal, instance-level privacy guarantees against MIAs. We refer to this approach as *with DP*, i.e., where AL and DP are jointly employed, and consider an alternative approach, namely, *without DP*, where DP is not employed. The goal is to evaluate the advantage of jointly leveraging AL and DP as a defense mechanism to mitigate CF-based shadow-based MIAs. Furthermore, we evaluate both model utility and the quality of the generated CFs, with the ultimate aim of analyzing how the proposed defense affects predictive performance and CF quality.

To initiate the AL process, we construct a labeled subset $\mathcal{S} \subset \mathcal{D}$ through random sampling. Initially, a small subset $\mathcal{S}_0 \subset \mathcal{D}$ is randomly selected to serve as the starting point for labeling. At iteration t, the current model f scores unlabeled data from $\mathcal{D} \setminus \mathcal{S}$. An acquisition function (e.g., entropy or uncertainty sampling) measures how *informative* each sample might be for improving the model. We select the most informative subset $\mathcal{S}_t$ from the pool using the least confidence metric. After adding the samples in $\mathcal{S}_t$, we incorporate them into the training set, i.e., $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_t$. The updated training set is then used in the next model retraining step $t+1$. These steps are repeated until a stopping condition is met, or a maximum number of iterations on a validation set is reached. *To integrate DP into the model training*, we update the model parameters following DP-SGD, where each weight in θ is updated during back-propagation step, where gradients computed using standard optimizers (e.g., SGD, Adam), are clipped to a fixed norm C, and Gaussian noise is chosen to satisfy a target (ϵ, δ)-DP guarantee. We consider the following scenarios:

- *Baseline*: Standard training without any additional techniques serving as a reference model for evaluating the benefits of DP and AL for preventing MIA.
- *Only-AL*: Incorporates AL to strategically select training samples with the aim of reducing the required training dataset size while maintaining predictive performance.
- *Only-DP*: Integrates DP into the training process of the model. It assesses the trade-offs between privacy and performance when DP is applied.
- *DP-Post-AL*: Applies AL first to strategically select training samples, and DP is then subsequently applied during the training phase. This setup assesses whether applying DP after AL can improve privacy preservation.
- *AL-Guided-DP*: Follows the same incremental training process as the default AL model, but with DP incorporated during each training step.

IV. Experimental Settings

Datasets. We consider two datasets for conducting our experiments, *EEG* [34] and *Indoor Location* (In-Location) [35]. We selected these two datasets for their suitability for training DNN models in terms of performance and practical relevance, and for the variation in the dimensionality of their input features. The EEG dataset comprises 14,980 data records, 14 numerical features, and 2 classes. The dataset contains brainwave and signal readings, with the objective of training a model to predict whether an eye is open or closed. In-Location comprises 20,000 data records and 529 features related to Wi-Fi signals, along with multiple labels, and is used to predict the building.

Model Architectures and Training Protocol. Preprocessing consists of removing missing values, removing outliers, and normalizing with a standard scaler. For both datasets, we split the data into 45% for training the target model, 45% for training shadow models, and 10% for validation. For the EEG *Baseline* model, we employed a DNN with 4 hidden layers containing 32, 16, 32, and 16 nodes, respectively. Each hidden layer was followed by a ReLU activation. The In-Location *Baseline* model was constructed with 7 layers, with 600, 700, 600, 300, 600, 300, and 128 nodes, respectively, followed by ReLU activation function, and an additional dropout layer was used to prevent overfitting. These models were optimized using the Adam optimizer to minimize the cross-entropy loss, with a learning rate of 0.01.

Defense Mechanisms. For scenarios where only DP was applied during training (i.e., *Only-DP*), we maintained the same model architecture as the default. Models were trained with privacy budgets (ε) of 1, 3, 5, and 10, with lower budgets corresponding to stronger privacy guarantees, and with δ = 1e-5, grad normalization of 1.5. In the *Only-AL* approach, the model during the first iteration is trained using 10% of the available training set. With each iteration, an additional 50 data records are added; the process is repeated for 50 iterations, and the model achieving the highest accuracy is selected. The training subset used to train the optimal model is then saved as the *Best AL Dataset*, on which the model achieves the best predictive performance. For model training in the *DP-Post-AL*, we use the *Best AL Dataset* and train models with DP.

Membership Inference Attack Setup. MIA is implemented by performing hyperparameter tuning on the number of shadow models (i.e., the DNNs) and attack models (XGBoost) for the shadow models' outputs on the both datasets. After hyperparameter tuning, the XGB attack model is configured with 150 estimators, a maximum depth of 15, and a learning rate of 0.01. The shadow dataset is collected by using a separate subset of the dataset to query the MLaaS. The results are presented as the average of 5 runs.

Counterfactual Explanation Setup. All CFs were obtained by using the Nearest Instance Counterfactual Explanations (NICE) [36] with the following criteria: Distance Metric of Heterogeneous Euclidean-Overlap Metric and Normalization of Min-Max. The algorithm was set to optimize proximity and sparsity for the *In-Location* dataset. CF quality was measured using common metrics such as average proximity and sparsity. Each data point from the testing set was explained individually, and its proximity and sparsity values were used to compute the average proximity and sparsity across all data points. We gain insight into how effectively the CF explainer can

TABLE I: Predictive performance (Accuracy, Precision, Recall) under privacy budgets ε for EEG and In-Location. *Baseline and Only-AL are trained without DP and therefore have no associated ε.*

Data	Method	Accuracy				Precision				Recall			
		ε				ε				ε			
		1	3	5	10	1	3	5	10	1	3	5	10
EEG	Baseline	96				96				96			
	Only-AL	95				95				95			
	Only-DP	82	87	88	90	82	87	88	90	82	87	88	90
	AL-Guided-DP	82	87	90	91	82	88	90	92	82	87	90	91
	DP-Post-AL	66	72	83	86	69	74	84	86	66	72	83	86
In-Location	Baseline	100				100				100			
	Only-AL	100				100				100			
	Only-DP	40	75	81	86	24	79	85	88	40	75	81	86
	AL-Guided-DP	58	57	56	91	46	56	67	92	58	57	56	91
	DP-Post-AL	47	45	52	53	23	29	50	54	47	45	52	53

produce *concise* (i.e., small-distance, few-feature changes) yet *impactful* (i.e., changing the predicted outcome) CFs under different privacy conditions.

Evaluation Metrics Setup. Relative to MIA, we propose two distinct metrics. Let $\mathcal{D}$ denote the full dataset with $|\mathcal{D}| = N$ records. Let $\mathcal{S} \subseteq \mathcal{D}$ be the set of records used to train a given model (e.g., depending on whether AL is applied). Let $\hat{\mathcal{S}} \subseteq \mathcal{D}$ be the set of records that an adversary predicts as *members* via MIA.

Micro-level defended record ratio (per model): At the micro level, defended records quantify the fraction of *true members* whose membership remains hidden, i.e., training points that the attacker *fails* to infer as members, where $\mathcal{D}_{\text{protected}} = \mathcal{S} \setminus \hat{\mathcal{S}}$ and true positive rate (TPR) among all actual members, the fraction the attacker correctly detects.

$$Micro = \frac{|\mathcal{D}_{\text{protected}}|}{|\mathcal{S}|} = 1 - \text{TPR} \tag{1}$$

Macro-level defended record ratio (overall exposure): At the macro level, we additionally account for records that were *never* used for training and are therefore automatically unexposed to MIA. It describes the fraction of all records whose membership is not successfully revealed, i.e., records that are either saved (never trained) or protected despite being trained, where $\mathcal{D}_{\text{saved}} = \mathcal{D} \setminus \mathcal{S}$.

$$Macro = \frac{|\mathcal{D}_{\text{saved}}| + |\mathcal{D}_{\text{protected}}|}{|\mathcal{D}|} = 1 - \frac{|\mathcal{S}|}{|\mathcal{D}|} \cdot \text{Recall}_{\text{member}}. \tag{2}$$

V. Experimental Results

We present the numerical results along four complementary dimensions. First, we evaluate predictive performance under the different defense mechanisms. Second, we assess MIA effectiveness with and without CFs to quantify the additional privacy risk introduced by CFs and the robustness of the defenses. Third, we report micro- and macro-level defended record ratios to provide a granular view of privacy exposure. Finally, we analyze the impact of the defense strategies on CF quality.

A. Model Predictive Performance

Table I summarizes the predictive performance of the different models across all approaches and across both datasets. Across both datasets, as expected, *Baseline*, which does not implement any privacy measure, achieves the highest performance across all metrics (e.g., accuracy, precision, and recall of 96% and 100% across EEG and In-Location, respectively), while *Only-AL*, which leverages AL to reduce the training set size while ensuring performance matches that of *Baseline*, closely follows (*Only-AL* achieves this level of performance using only a subset of the dataset).

When DP is incorporated as part of the defense mechanism, namely in *Only-DP*, *AL-Guided-DP*, and *DP-Post-AL*, a consistent pattern emerges across both datasets. Stronger privacy guarantees (lower ε) lead to a noticeable degradation in performance, whereas performance steadily improves as ε increases, corresponding to weaker privacy constraints.

In the EEG dataset, *AL-Guided-DP* and *Only-DP* exhibit comparable performance and consistently outperform *DP-Post-AL*, which yields the lowest scores. Specifically, *AL-Guided-DP* achieves accuracy between 82% and 91%, with precision ranging from 82% to 92%, and recall following a similar pattern. *Only-DP* attains accuracy between 82% and 90% across different ε values. For In-Location, DP exerts a stronger impact on performance than in EEG, although the overall trend remains similar. *Only-DP* and *AL-Guided-DP* alternate in achieving the best predictive results, with comparable performance levels. Importantly, *AL-Guided-DP* attains this performance while relying on fewer training samples (since it employs AL), which is beneficial from a privacy perspective as it reduces potential exposure to MIAs.

Regarding the subset size selected by AL, for EEG, *Only-AL* (and *DP-Post-AL*) uses 77.45% of the data. *AL-Guided-DP* selects 78.97%, 74.48%, 56.49%, and 66.98% for $\varepsilon = 1, 3, 5,$ and 10, respectively. For In-Location, *Only-AL* uses only 10% of the data, while *DP-Post-AL* requires 29.13%, 21%, and 40.35% for $\varepsilon = 1, 3,$ and 5. Overall, AL can substantially reduce the required training data; however, when combined with DP, larger subsets are often selected, reflecting the additional uncertainty introduced by DP noise.

Takeaway 1: AL preserves utility, matching baseline performance while using fewer training samples. Introducing DP leads to a utility loss at low ε. While *AL-Guided-DP* does not substantially outperform *Only-DP*, it achieves comparable performance with fewer samples. In contrast, *DP-Post-AL* consistently yields the lowest performance across both datasets.

B. Membership Inference Attack

Figure 2 summarizes MIA performance across all approaches and privacy budgets ε of 1,3,5 and 10 (lower values indicate stronger privacy protection). Results show a clear trend across all cases: *leveraging CFs to perform MIA allows for substantially amplifying attack success*. For instance, in the

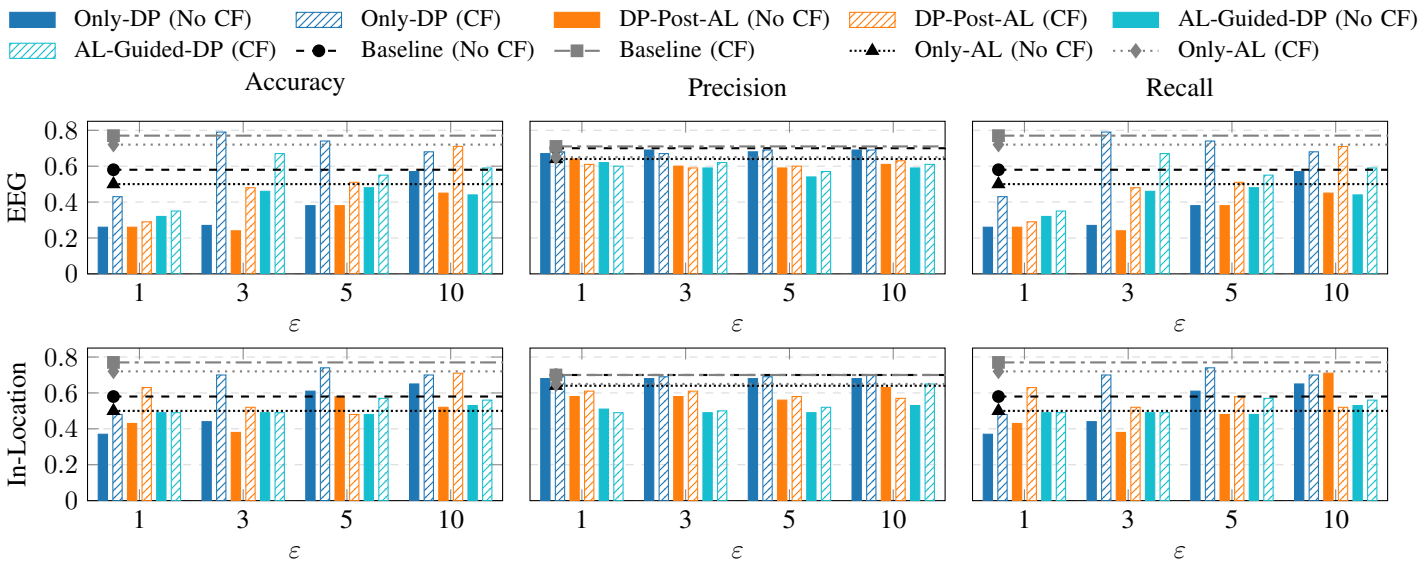

Fig. 2: Accuracy, precision, and recall across baselines, privacy budgets ε with and without using CFs for MIA.

EEG dataset, *Baseline (No CF)* attains 58% accuracy/recall and 70% precision while *Baseline (CF)* attains 77% of accuracy/recall, and 71% of precision. Another example is *Only-AL (No CF)*, accuracy/recall increases from 50% to 72% (precision 64-65%) when leveraging CFs (*Only AL (CF)*. When employing DP, the impact of CFs on MIA success becomes even more evident. Comparing *Only-DP (no CF)* to *Only-DP (CF)*, accuracy/recall increases from 26% to 43% at $\varepsilon = 1$, from 27% to 79% at $\varepsilon = 3$, from 38% to 74% at $\varepsilon = 5$, and from 57% to 68% at $\varepsilon = 10$ when CFs are leveraged, while precision remains essentially close (67–69%). Results in In-Location show the same trend: leveraging CFs consistently strengthens MIA. For example, *Only-DP*, leveraging CFs raises accuracy/recall from $37, 44, 61, 65\%$ to $48, 70, 74, 70\%$ at $\varepsilon = 1, 3, 5, 10$. For *DP-Post-AL*, accuracy increases from $43, 38, 58, 52\%$ to $63, 52, 48, 71\%$, except at $\varepsilon = 5$. *AL-Guided-DP* shows little change at $\varepsilon = 1, 3$, a moderate gain at $\varepsilon = 5$, and at $\varepsilon = 10$ its largest precision increase (53% to 65%) with a slight accuracy rise (53% to 56%). *Takeaway 2:* While DP introduces a utility–privacy trade-off (Takeaway 1), leveraging CFs significantly weakens privacy protection across all settings a consistently amplifies MIA performance, often dramatically increasing attack success, even under DP. In several cases, CFs nearly double accuracy/recall at moderate privacy budgets, demonstrating that explanations can substantially undermine the privacy gains achieved by DP alone.

We now quantify the impact of ε on the defense against MIA. As expected, a larger ε weakens privacy: for example, on EEG, *Only-DP (No CF)* accuracy/recall range from 26% to 57% and from 43% to 68% *Only-DP (CF)*. Comparable monotonic trends appear in *DP-Post-AL* and *AL-Guided-DP*. We move to comparing *Only-AL to DP-Post-AL* in order to better understand the effect of employing DP after AL. For EEG, *Only-AL (No CF* has an accuracy and recall $\approx 50\%$, which reaches 72% when CFs are exposed. In contrast, *DP-Post-AL* starts from lower attack success at small privacy budgets (26% and 34% at ε 1 and 3), indicating stronger protection than *Only-AL*. However, CFs significantly erode this advantage: accuracy rises to 29% and 48% at $\varepsilon = 1$ and 3, and at larger ε of 10 reaching 71%. For In-Location, similar trends are observed.

Takeaway 3: Consistent with the utility–privacy trade-off observed in Takeaway 1, increasing ε systematically weakens protection and enables stronger MIAs. However, beyond this expected trend, releasing CFs introduces an additional attack surface that substantially boosts member detection, primarily by increasing recall. Even when DP reduces attack success at low privacy budgets, this advantage is significantly diminished once CFs are exposed. While combining AL with DP can partially mitigate attack success, this comes at the cost of retaining more training data.

C. Defense at Micro- and Macro-Level

In this part we focus our analysis on the two approaches, *DP-Post-AL* and *AL-Guided-DP*, and examine their behavior through the micro and macro defended record ratios. The micro metric captures the fraction of *true training records* whose membership remains hidden, directly quantifying how well individual members are protected. In contrast, the macro metric reflects overall dataset exposure by additionally accounting for records that were never used for training and are therefore automatically unexposed to MIA. Consequently, macro values can remain high even when the protection of actual training points deteriorates.

Across all settings, fig. 3 shows that leveraging CFs consistently reduces the proportion of defended points at both levels, with the effect being particularly pronounced at the micro level. On EEG, under *DP-Post-AL*, micro protection drops from 68% to 44% at $\varepsilon = 5$, and from 56% to 9% at $\varepsilon = 10$ when CFs are leveraged. A similar trend is observed for *AL-Guided-DP*, where micro defended rates decrease from 54%

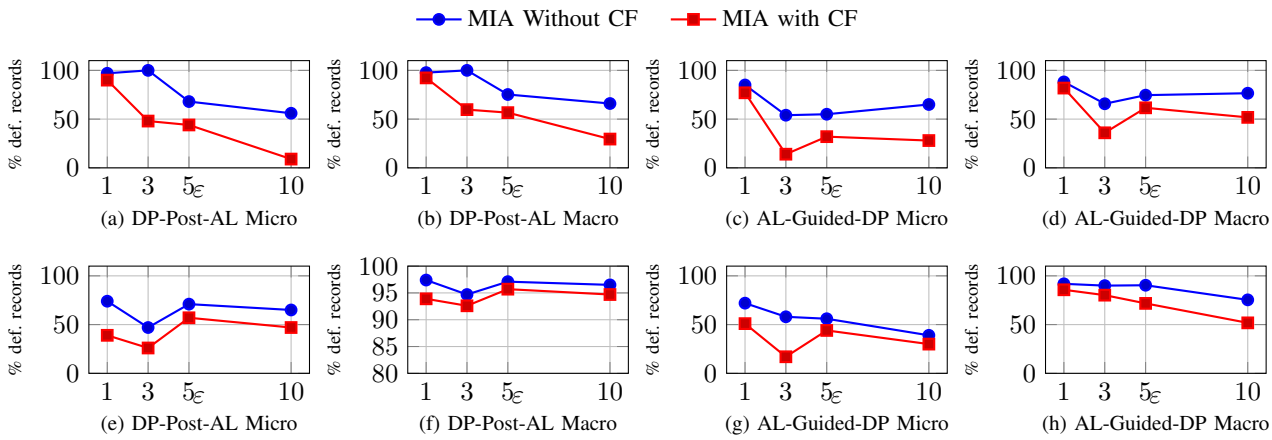

Fig. 3: Percentage of micro & macro defended records under *DP-Post-AL* & *AL-Guided-DP*, with & without CFs. Top row: EEG; bottom row: In-Location.

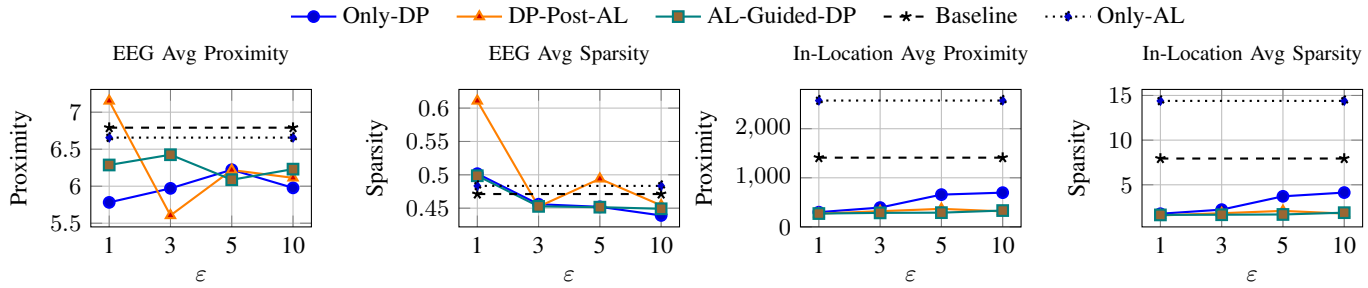

Fig. 4: Sparsity and proximity of extracted CFs for EEG and In-Location across differential privacy budgets (ε).

to 14% at $\varepsilon = 3$. These results indicate that CFs substantially increase correct member inference, exposing training data.

At the macro level, defended ratios also decrease when CFs are used, but the reduction is often less because macro protection includes automatically saved (non-trained) records. As a result, an approach may appear robust at the macro level while failing to adequately protect actual training members. For instance, in EEG at $\varepsilon = 3$, *AL-Guided-DP* decreases from 65.74% to 35.95% at the macro level, whereas the corresponding micro-level performance is considerably stronger. For EEG, *DP-Post-AL* achieves higher macro defended ratios than *AL-Guided-DP* at small ε, consistent with stronger privacy guarantees at low budgets, although this advantage is not uniform at larger ε. Similar patterns emerge for In-Location, in which CFs consistently erode micro-protection across ε.

Takeaway 4: Beyond amplifying MIA accuracy (Takeaway 2) and weakening protection as ε increases (Takeaway 3), CFs critically erode *micro-level* protection, directly exposing true training members. While macro-level defended ratios may remain relatively high due to automatically saved records, the micro analysis reveals a substantial loss of protection for most of the actual training data. These results indicate that aggregate exposure metrics can mask individual-level vulnerability, and that releasing CFs can significantly undermine the practical privacy guarantees of DP-based approaches.

D. CF Quality Analysis

Figure 4 reports proximity and sparsity across scenarios. For EEG, proximity remains stable under DP (5.78–6.22 for *Only-DP*; 5.60–7.15 for *DP-Post-AL*), indicating that DP does not substantially degrade CF quality. Notably, *AL-Guided-DP* exhibits slightly improved proximity compared to *Only-AL*, which may be attributed to its use of a larger training subset. By incorporating more data points during training, *AL-Guided-DP* may learn more stable decision boundaries, resulting in CFs that require smaller input changes. For In-Location, proximity values are generally higher than the *Baseline*, reflecting the dataset's higher dimensionality. Sparsity follows a similar pattern: under tight privacy budgets (e.g., $\varepsilon = 1$), DP slightly reduces sparsity, indicating stronger noise effects. In summary, proximity remains largely stable across privacy levels, whereas sparsity is mildly affected and exhibits greater sensitivity in high-dimensional settings.

Building on our analysis, *Takeaway 5:* While DP introduces a utility–privacy trade-off (Takeaway 1) and CFs substantially amplify MIA success (Takeaways 2–4), the impact of DP on CF quality remains limited. Proximity remains stable across privacy budgets, and sparsity is only mildly affected, particularly in high-dimensional datasets.

These findings reveal a fundamental tension: explanations can significantly weaken privacy guarantees, yet privacy-preserving mechanisms can be applied without substantially degrading explanation quality. This creates a critical design challenge for MLaaS systems, which must carefully balance predictive utility, privacy robustness, and explanation fidelity. Beyond technical performance metrics, future work should examine the practical usefulness of CFs under privacy-preserving

settings through empirical user studies. In particular, it is important to assess whether explanations that remain stable under DP also retain their interpretability and decision-support value for end users in high-stakes domains.

VI. Conclusion

In this paper, we investigate how counterfactual explanations (CFs) expand the attack surface of Machine Learning as a Service systems by enabling highly effective shadow-based membership inference attacks. We propose a defense framework that integrates Differential Privacy (DP) and Active Learning, and systematically analyze the resulting interplay between privacy protection, predictive performance, and explanation quality. Our results demonstrate that while DP can mitigate membership leakage, its protection can be substantially weakened when CFs are exposed. At the same time, we show that privacy-preserving training does not necessarily degrade explanation quality, revealing a fundamental tension between transparency and privacy resilience.

References

[1] T. T. Nguyen, T. T. Huynh, Z. Ren, T. T. Nguyen, P. L. Nguyen, H. Yin, and Q. V. H. Nguyen, "Privacy-preserving explainable ai: a survey," *Science China Information Sciences*, vol. 68, no. 1, p. 111101, 2025.

[2] F. Ezzeddine, "Privacy implications of explainable ai in data-driven systems," 2024.

[3] 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.

[4] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[5] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[6] 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.

[7] 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.

[8] A. C. Oksuz, A. Halimi, and E. Ayday, "Autolycus: Exploiting explainable artificial intelligence (xai) for model extraction attacks against interpretable models," *Proceedings on Privacy Enhancing Technologies*, 2024.

[9] M. Veale, R. Binns, and L. Edwards, "Algorithms that remember: model inversion attacks and data protection law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180083, 2018.

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

[11] R. Mochaourab, S. Sinha, S. Greenstein, and P. Papapetrou, "Robust counterfactual explanations for privacy-preserving svm," in *International Conference on Machine Learning (ICML 2021), Workshop on Socially Responsible Machine Learning*, 2021.

[12] F. Ezzeddine, O. Ayoub, and S. Giordano, "Knowledge distillation-based model extraction attack using private counterfactual explanations," *arXiv preprint arXiv:2404.03348*, 2024.

[13] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5558–5567.

[14] R. Munro, *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning, 2021.

[15] N. Sachdeva and J. McAuley, "Data distillation: A survey," *arXiv preprint arXiv:2301.04272*, 2023.

[16] S. Lei and D. Tao, "A comprehensive survey of dataset distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 17–32, 2023.

[17] T. Dong, B. Zhao, and L. Lyu, "Privacy for free: How does dataset condensation help privacy?" in *International Conference on Machine Learning*. PMLR, 2022, pp. 5378–5396.

[18] T. Zheng and B. Li, "Differentially private dataset condensation," 2023.

[19] J. Ferry, U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala, "Taming the triangle: On the interplays between fairness, interpretability, and privacy in machine learning," *Computational Intelligence*, vol. 41, no. 4, p. e70113, 2025.

[20] H. Liu, Y. Wu, Z. Yu, and N. Zhang, "Please tell me more: Privacy impact of explainability through the lens of membership inference attack," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 4791–4809.

[21] 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.

[22] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 634–646.

[23] 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.

[24] J. Li, N. Li, and B. Ribeiro, "Membership inference attacks and defenses in classification models," in *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, 2021, pp. 5–16.

[25] 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.

[26] 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.

[27] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model." *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.

[28] D. Ye, S. Shen, T. Zhu, B. Liu, and W. Zhou, "One parameter defense—defending against data inference attacks via differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1466–1480, 2022.

[29] 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.

[30] R. Naidu, A. Priyanshu, A. Kumar, S. Kotti, H. Wang, and F. Mireshghallah, "When differential privacy meets interpretability: A case study," *arXiv preprint arXiv:2106.13203*, 2021.

[31] F. Ezzeddine, M. Saad, O. Ayoub, D. Andreoletti, M. Gjoreski, I. Sbeity, M. Langheinrich, and S. Giordano, "Differential privacy for anomaly detection: Analyzing the trade-off between privacy and explainability," in *World Conference on Explainable Artificial Intelligence*. Springer, 2024, pp. 294–318.

[32] S. Saifullah, D. Mercier, A. Lucieri, A. Dengel, and S. Ahmed, "The privacy-explainability trade-off: unraveling the impacts of differential privacy and federated learning on attribution methods," *Frontiers in Artificial Intelligence*, vol. 7, p. 1236947, 2024.

[33] H. Nori, R. Caruana, Z. Bu, J. H. Shen, and J. Kulkarni, "Accuracy, interpretability, and differential privacy via explainable boosting," in *International conference on machine learning*. PMLR, 2021, pp. 8227–8237.

[34] O. Roesler, "Eeg eye state [dataset]," *UCI Machine Learning Repository*, 2013.

[35] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems," in *2014 international conference on indoor positioning and indoor navigation (IPIN)*. IEEE, 2014, pp. 261–270.

[36] D. M. D. Brughmans, P. Leyman, "Nice: an algorithm for nearest instance counterfactual explanations," *Data Mining and Knowledge Discovery*, 2023.