Knowledge Distillation-Based Model Extraction Attack using GAN-based Private Counterfactual Explanations

Fatima Ezzeddine
Faculty of Informatics
Università della Svizzera italiana
fatima.ezzeddine@usi.ch

Omran Ayoub
Department of Innovative Technologies
University of Applied Sciences and
Arts of Southern Switzerland
omran.ayoub@supsi.ch

Silvia Giordano
Department of Innovative Technologies
University of Applied Sciences and Arts
of Southern Switzerland
silvia.giordano@supsi.ch

Abstract

In recent years, there has been a notable increase in the deployment of machine learning (ML) models as services (MLaaS) across diverse production software applications. In parallel, explainable AI (XAI) continues to evolve, addressing the necessity for transparency and trustworthiness in ML models. XAI techniques aim to enhance the transparency of ML models by providing insights, in terms of *model's explanations*, into their decision-making process. Simultaneously, some MLaaS platforms now offer explanations alongside the ML prediction outputs. This setup has elevated concerns regarding vulnerabilities in MLaaS, particularly in relation to privacy leakage attacks such as model extraction attacks (MEA). This is due to the fact that explanations can unveil insights about the inner workings of the model which could be exploited by malicious users. In this work, we focus on investigating how model explanations, particularly counterfactual explanations (CFs), can be exploited for performing MEA within the MLaaS platform. We also delve into assessing the effectiveness of incorporating differential privacy (DP) as a mitigation strategy. To this end, we first propose a novel approach for MEA based on Knowledge Distillation (KD) to enhance the efficiency of extracting a substitute model of a target model exploiting CFs, without any knowledge about the training data distribution by the attacker. Then, we advise an approach for training CF generators incorporating DP to generate private CFs. We conduct thorough experimental evaluations on real-world datasets and demonstrate that our proposed KD-based MEA can yield a high-fidelity substitute model with a reduced number of queries with respect to baseline approaches. Furthermore, our findings reveal that including a privacy layer can allow mitigating the MEA. However, on the account of the quality of CFs, impacts the performance of the explanations.

1 Introduction

Recent years have witnessed a growing trend in employing deep neural networks (DNNs) as learning algorithms for machine learning-based applications. In particular, DNNs have gained substantial popularity due to their remarkable success in diverse domains [52]. However, the complexity of their training process, which is resource-intensive, involves the acquisition of data and requires substantial

2nd Workshop on Regulatable ML at NeurIPS 2024.

computational power [30], hinders their widespread adoption and accessibility. An approach to addressing this challenge is by offering machine learning (ML) models as a service through dedicated ML as a Service (MLaaS) platforms [36, 53]. MLaaS platforms allow to streamline the accessibility of such complex models by hosting and training ML models on the cloud, allowing third-party practitioners to access these pre-trained models through Application Programming Interfaces (APIs) [60], while the private dataset used to train the model is inaccessible.

Numerous platforms offer MLaaS such as Google Cloud, Microsoft, IBM, and Amazon Web Services [1, 6, 7, 8, 9, 10]. The API's input-output format for these services is publicly accessible, meaning that users have knowledge of the format of input data required by the service (by the ML model), and can interpret its outputs. This opens the door for malicious users (or, more precisely, attackers) to attempt to perform privacy-breaching attacks such as membership inference attacks (MIA) [54], model inversion attacks (MINA) [59], and model extraction attacks (MEA) [57]. For instance, in MIA, an attacker aims to infer the presence of a specific individual instance in the private training set. In MINA, the attacker attempts to reconstruct the training set. In MEA, the attacker attempts to extract the model by training a substitute model using data acquired through repetitive queries to the service provider's confidential model. Several mitigation strategies for these security- and privacy-breaching attacks are already employed [17]. However, these strategies are faced with new challenges as emerging transparency requirements demand explanations for ML model decisions. In fact, MLaaS platforms are shifting towards providing explanations alongside the predictions made by the deployed ML models [1, 7, 8, 10]. The need to provide users with such model explanations is linked to providing users with a transparent decision-making process of data-driven systems [40]. The urge to enhance transparency through model explanations aligns, on one hand, with the aim of meeting recent regulatory (and compliance) demands, and on the other hand, with the aim of providing users with actionable insights [38, 40, 56]. Such explanations are extracted using explainable AI (XAI) techniques, which aim to provide insight into how a model arrived at its decision [21, 50]. In this context, explanations may inadvertently provide attackers with insights to enhance their attacks, posing a fresh obstacle to existing security and privacy measures [22, 23, 47].

In this work, we focus on investigating how example-based explanations, specifically counterfactual (CF) explanations, can be exploited to perform MEA within MLaaS scenarios. We precisely focus on CFs due to the unique insights they provide, as they reveal how to minimally alter original data instances to achieve a different model's outcome, allowing users to interpret the dynamics underlying the model's prediction shift from the decision boundary [15]. However, due to the proximity of CFs to the decision boundary and their representativeness of the training data, CFs possess a dual nature, serving not only as interpretative tools but also as potential assets for enhancing attacks as they provide additional insights and knowledge for potentially exploiting the vulnerabilities of the model under attack. Specifically, we propose a novel approach based on the concept of knowledge distillation (KD) to perform high-fidelity MEA with CFs as the representation of data, and derive a threat model that have similar functionality of the original model being targeted [1]. The unique capabilities of KD allow us to handle the problem not just as a standard classification task but as an estimation of the output probability distribution. Moreover, we propose a mitigation approach that employs differential privacy (DP) within the CF generator training pipeline, aiming to reduce the risk associated with providing CF explanations. Our analysis examines the attack performance of our proposed method compared to baselines. Additionally, we assess the impact of DP on the quality of CF and the performance of the MEA. The contribution of this paper is summarized as follows:

- **KD-Based MEA with CFs**: This research investigates potential vulnerabilities in MLaaS, with a specific focus on MEA facilitated by KD as extraction techniques and CFs as representatives of the training set. The research simulates an adversarial scenario wherein an attacker employs our proposed KD techniques to extract a substitute for a target model.

- **Private Counterfactual Explanations**: Recognizing the importance of preserving the privacy of training data, the paper introduces the concept of DP within the GAN-generated CF explanation pipeline. This contribution aims to generate CFs that deviate from the statistical properties of the confidential dataset, offering a layer of protection against potential privacy breaches.

[1] Link to code: https://github.com/FatimaEzzedinee/Knowledge-Distillation-Based-Model-Extraction-Attack-using-GAN-based-Private-CFs

2 Related work

Several works have proposed strategies for performing MEA for classification. Authors in [57] perform successful MEA on different ML models like decision trees, Support Vector Machines, and DNNs by using equation-solving and path-finding algorithms and learning theory. In [45] authors perform MEA on DNNs using active learning with unannotated public data. In [28], authors perform MEA in natural language processing on bert-based APIs, in which they explore transfer learning for MEA. Other works have focused on proposing querying strategies for MLaaS to effectively query the target model and extract accurate insights in addition to performing MEA [26, 42, 44]. In [44] authors train a knockoff network with queried image predictions and propose a reinforcement learning approach that additionally improves query sample efficiency in certain settings and provides performance gains. In [26] authors propose a new method that generates synthetic queries and optimizes training hyperparameters, and then propose a method to detect generic and effective detection of DNN MEA. In [42] authors investigate the type and amount of internal information about the black-box model that can be extracted from querying the deployed model, such as the architecture, optimization procedure, or training data. Authors in [24] improve the query efficiency of attacks by designing learning-based methods and focusing on the real accuracy of the extracted model. Also [29], discuss the tension between the right to explanation and the right to be forgotten for privacy concerns.

Other works have investigated how to reveal insights of ML models using the explanations provided by the MLaaS [12, 32, 43, 55, 61]. In [43] authors investigate how Local Interpretable Model-agnostic Explanations (LIME), an XAI framework, can be exploited to infer the decision boundaries by sending adaptive queries to generate new data samples that are close to the decision boundaries. In [33] authors analyze how gradient-based explanations reveal the decision boundary of a target model by modeling a data-free MEA against a gradient-based XAI, in addition to exploiting a generative model to reduce the number of queries. [61] proposed a methodology to perform MEA by minimizing task-classification loss and task-explanation loss. Similar to our work, other efforts have investigated how CFs can be exploited in MEA. In [12] authors model an attack that relies on both the predictions and the CFs of the target model to directly train an attack model. The authors in [60] model a strategy to reveal the decision boundary of a target model and then perform MEA by considering the CF of the CF as pairs of training samples to directly train the extracted model.

Similar to these works, our work aligns with exploring how CF explanations can be exploited for performing MEA. However, our contribution lies in modeling a novel approach to MEAs that exploits more effectively the insights that CFs carry with respect to other approaches that directly train a substitute model on extracted CFs. Moreover, in contrast to previous studies [12, 60], our work presumes that the attacker possesses no prior knowledge of the training set, leading them to query the model with zero knowledge, and propose a novel approach for performing the MEA.

In terms of mitigation strategies, few works have focused on proposing methodologies to generate explanations while mitigating the revelation of sensitive insights about the decision boundary, the training set, or the model architectures. For instance, authors in [46] propose employing DP algorithms to construct feature-based model explanations, while authors in [62] propose an approach to generate differentially private CFs via functional mechanisms. [48] proposes an approach for generating recourse paths leveraging differentially private clustering and demonstrates that constructing a graph on the cluster centers provides private recourse paths as CFs.

Also, our work also exploits DP, however, by integrating it into the CF generation process, and eliminates the necessity for a distinct method to generate private CFs. We leverage the strength of generative adversarial networks (GANs)-based generators in producing high-quality, yet private, CFs. To the best of our knowledge, our work is the first to adapt DP within CF generative-based methods. We specifically focus on GANs-based CF generators due to their ability to generate high-quality CFs compared to other CF generation methods. We show the robustness of the DP-based proposed method against MEA, and we provide a detailed analysis that sheds light on the impact of CFs in exposing the distribution of the training set through well-defined metrics.

3 Scenario and problem formulation

A DNN model f_θ trained on a dataset $\mathcal{D} : \mathcal{X} \in \mathbb{R}^d \rightarrow \mathcal{Y} \in \mathbb{R}$ takes a numerical input query $x \in \mathcal{X}$ and predicts the output $y \in \mathcal{Y}$. f_θ a trained model with θ be the weight matrix, b be the bias vector.

The output of the DNN can be computed as $f_\theta = \theta \cdot X + b$. The output of the network y are confidence scores, which are expressed in terms of probabilities, essentially indicating how confident the ML model is about each possible class or category in its training data. The CF explanation method, which trains a CounterGAN generator E: $f_\theta \times \mathcal{X} \rightarrow \mathcal{X}$ generates a perturbed instance $c \in \mathcal{X}$ for the input instance x such that $f_\theta(c)$ has a different output while minimizing an objective function g. Formally, searching for a CF can be framed as $argmin\ g(x, c)\ s.t\ f_\theta(c) \neq f_\theta(x)$. Where $g : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{R}_+$ is a cost metric measuring the changes between the input x and c, and y is the desirable target that is different from the original prediction f_θ.

Figure 1 depicts the steps of the scenario. First, a user sends a query, which includes input data describing a data record x for which the user aims to receive the prediction. Once the MLaaS API receives the user query, it will pass it to the ML service, which performs the prediction $f_\theta(x)$ and generates a CF explanation $c = E(x)$ where $f_\theta(c)$ has a different prediction and finally returns it along with its corresponding output to the user.

Within the scope of this research, our primary emphasis is placed on a specific type of attack known as the high-fidelity MEA, specifically directed toward CFs. The objective of this attack is to derive a model threat_model t_Υ that closely resembles, in terms of functionality, the original model being targeted.The MEA is formulated as follows: for a set of queries $\mathcal{Q}$ and a set of corresponding CF explanations $\mathcal{C}$, an attacker trains a threat model t_Υ that performs equivalently on an evaluation set $\mathcal{T}$. The goal is to maximize an agreement function between f_θ and t_Υ as shown in Eq. 2, with a minimal number of queries sent to the API. The agreement metric will be detailed later in this paper in Sec. 5.

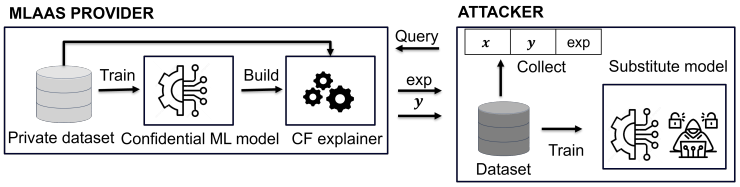


Figure 1: Scenario of MEA where MLaaS provides explanations alongside the prediction

4 Methodology

4.1 Knowledge distillation-based model extraction attack

First, an owner of a private dataset trains a DNN classifier (Step 1 in Fig. 2), and a CounterGAN CF explainer (Step 2A, 2B in Fig. 2) and deploys it on an MLaaS platform. As an attacker, specifically, we query the model with random queries [2] and collect the query output of the target model, and the CFs explaining the output of the target model (Step 3 in Fig. 2). After successfully collecting a set of CFs (Step 4 in Fig. 2), the attacker employs our KD-based MEA approach on the collected dataset as input data to train t_Υ. The rationale for selecting KD as the extraction method lies in its unique capability to address the problem not only as a standard supervised classification task but also in its proficiency in precisely imitating the probability distribution. Specifically, we propose to train t_Υ by minimizing the loss constituted by the threat model classification loss in addition to the distillation loss. To emphasize the importance of mimicking the output probabilistic distribution from the target model to the threat model, we set the distillation loss to the Jensen-Shannon (JS) (1) metric. The choice of JS is to tackle issues related to divergence and ensure a symmetric output, so we suppose that the attacker replaces the KL divergence metric with JS. Afterward, the attacker trains the threat model by minimizing the loss of KD (Step 5 in Fig. 2) until the convergence of the agreement on a separate validation set.

$$\text{JS}(P\|Q) = \frac{1}{2}\text{KL}\left(P \middle\| \frac{P+Q}{2}\right) + \frac{1}{2}\text{KL}\left(Q \middle\| \frac{P+Q}{2}\right) \tag{1}$$

4.2 Private CFs generation with DP

We are now aligned with the service provider and are presenting a strategy to mitigate MEA by avoiding providing Private CFs. Our goal is to examine the effects of incorporating DP into the CF

[2]The attacker does not have previous knowledge of the training set.

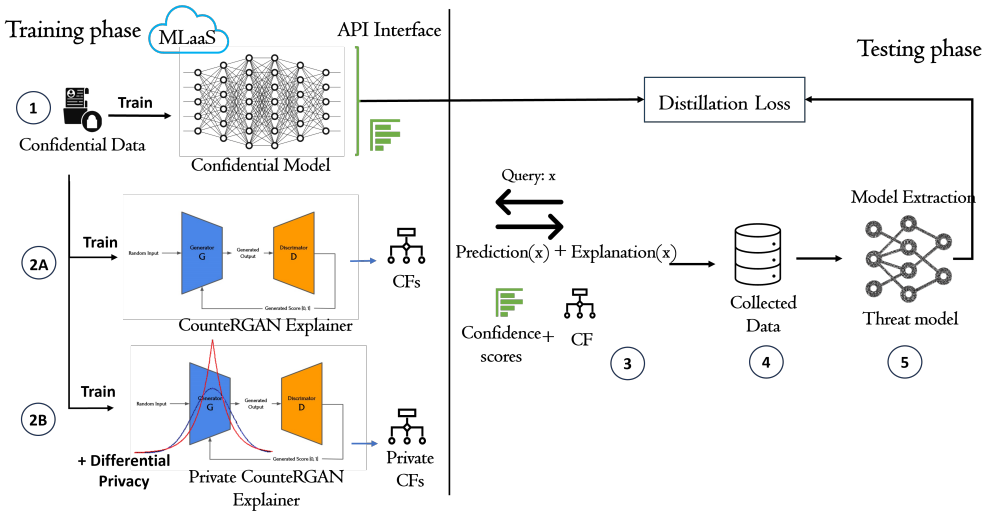


Figure 2: Methodology Framework, describing the KD-based MEA, in the deployment phase, in addition to the CF generation phase.

generation process on MEA performance and the quality of the provided CFs considering metrics such as actionability, prediction gain, and realism, which are explained later in sec 5.5. Additionally, we aim to demonstrate how private CFs align with the distribution of the training set. We propose an approach to integrating DP into the CF generation process and generating private CFs. Our methodology does not memorize or expose statistical information about the training set. We propose to incorporate DP into the generator of the CounteRGAN. The primary objective is to prevent the generation of CFs that closely resemble the private training data, thereby reducing the resemblance of CFs with the training data.

To this end, we inject DP into the generator during the optimization process. More specifically, we employ the Adam differential private optimizer (DP Adam). The process of DP Adam often involves multiple iterations of the training process to achieve privacy guarantees. We add noise repeatedly over several rounds, so the overall privacy guarantees are strengthened. To this end, we inject noise into the gradients of the generator in addition to a step of gradient clipping. The process involves clipping the gradient for a random subset of examples, clipping the norm of each gradient, computing the average, adding calibrated noise, and finally performing the traditional back-propagation step that updates the weights. Note that the gradient clipping step is essential as it controls the amount of noise added and prevents large updates that could reveal too much information about individual data points. The key assurance provided by this approach is that the generator ensures that it does not memorize or reveal details of the training data and therefore does not generate CFs that are very similar to the training points.

5 Experimental settings

5.1 Baseline scenarios

We compare our proposed approach for performing MEA, KD-based MEA, to a baseline approach, presented in [12], and referred to as *Direct Train*, which trains a threat model as a standard supervised classification task.

We examine three scenarios pertaining to the MLaaS provider's inclusion or exclusion of CF explanations alongside the predictions of the ML model, 1) *No CF:* In this scenario, the MLaaS does not provide the user with any CFs. 2) *CF:* In this case, the MLaaS provides CFs suing counterGAN. *3) Private CF:* The MLaaS offers private CFs by implementing our proposed approach for generating private CFs, discussed in Sec. 4.

Applying each MEA approach to the three scenarios results in the following six scenarios: 1) *KD-based Private CF.* We employ our proposed KD-based approach on CFs extracted from the

differentially private generator, and the focus shifts to examining the potential effect of integrating DP in the CF generation. 2) *KD-based CF.* We employ our proposed KD-based approach, trained with CFs, to understand the influence of KD when applied alongside CF. 3) *KD-based No CF.* We employ our proposed KD-based approach on the randomly generated data points, and the focus shifts to examining the potential enhancement when KD is utilized independently of CFs and analyzing the power of CFs alongside. 4) *Direct No CF.* The MEA is performed following the baseline [12]. We train a model directly on the randomly generated data points, and the focus shifts to examining the potential of direct queries and analyzing the power of KD and CFs. 5) *Direct CF.* We train a model directly on CFs, this dimension is dedicated to evaluating the standalone power of CFs without the incorporation of KD, providing insights into the capabilities of CFs separately. 6) *Direct Private CF.* We train a model directly on CFs extracted from the differentially private generator, the focus shifts to examining the potential of direct queries and to analyzing the power of KD and CFs alongside.

5.2 Target model

We adhere to consistent training procedures for the target model (to be deployed and queried by the API). We train a DNN comprising of 16 hidden layers to create a complex model. The layer configurations consist of 64, 32, 16, 32, 64, 128, 64, 32, 128, 64, 128, 64, 128, 64, 32, and 16 neurons in each layer, respectively. The activation function is set to gelu [20] for all layers, while the output layer employs softmax to produce model outputs as confidence score probabilities. We specify the optimizer as the Adam optimizer to minimize the cross-entropy loss function, with default parameter initialization provided by Keras. Each target model is trained using 80% of the training dataset [3].

5.3 Threat model

We specify consistent training procedures for the threat model, with the assumption that the attackers do not know the target model architecture but can build a DNN with a standard architecture. The attacker aim to train either by using our proposed approach of KD-based MEA or directly on the data points a DNN with 3 hidden layers of 16, 32, and 64 neurons consecutively, with an activation function as the relu, and an output layer with a softmax activation function. In the scenario we are considering, we assume that the attacker does not possess any knowledge about the distribution of the training data used for training the target model. Consequently, the attacker generates 1000 random data points for each dataset. The data point values are randomly generated, specifying the values of a range between -3 and 3 for each feature. We specify -3 and 3 as ranges for random values to have as randomized as possible data points that are not similar to the training datasets. For the training with KD-based MEA, we vary α between 0 and 0.5 and report the highest agreement. We made sure that the results were reproducible by fixing the seed and the layer initialization with *LeCun Initializer*. We report the results of the average of 10 runs with randomly chosen subsets for each experiment.

5.4 Counterfactual generator

To generate CFs using CountRGAN, we train a discriminator structured with three layers comprising 32, 16, and 1 neurons consecutively. Each hidden layer is followed by a dropout layer with a factor of 0.2. We use a relu activation function for the hidden layers and a sigmoid activation function for the output layer. The generator, which takes a datapoint as input and produces the corresponding CF, is trained with three hidden layers comprising 64, 48, and 32 neurons, using relu activation for the hidden layers and a linear activation function for the output layer. We specify the optimizer as Adam. For the DP generator, we change the optimizer to be an optimizer that supports DP. We integrate the DP Adam optimizer from the TensorFlow privacy library. To guarantee a good privacy budget, we specify the parameters of the DP Adam optimizer as l2_norm_clip to 1, the noise_multiplier to 3.

5.5 Evaluation metrics

The evaluation metrics comprise two sets. The first set measures the effectiveness of MEA while the second encompasses a pair of metrics to quantify the quality of the CFs.

Effectiveness of MEA: A commonly employed metric for assessing the effectiveness of MEA is the *Agreement* measure. *Agreement* measures the degree of alignment between two models, i.e., the

[3] Experiments were computed with a machine of intel Core i7, a GPU of RTX 3070, and 8 GB of RAM.

similarity in the predictions between the two different ML models. In the context of MEA methods, agreements assess the output similarity between the predictions of an extracted model t_Υ and those of the target model f_θ for a given set of data records (see Eqn. 2). Since our goal is to replicate the behavior of a target model with an extracted model, a higher agreement means a more successful MEA, and hence, a more effective strategy for MEAs.

$$agreement(f_\theta, t_\Upsilon) = \sum_{x_i \in \mathcal{T}} 1_{f_\theta = t_\Upsilon} \tag{2}$$

Success of explainer and quality of CFs: To measure the quality of extracted CFs, and to analyze the explainer prediction classes shift we employ three commonly used metrics, namely, *Prediction Gain*, *Realism*, and *Actionability*.

The *Prediction Gain* quantifies how much the probability of the CF explanation for a specific target class t changes in comparison to the original data point as provided by the classifier for a given target class t. In other words, it allows us to quantify to what extent the explainer was capable of shifting the prediction of the classifier using the extracted CFs. In our case, since the classifier's score is in terms of probabilities, the prediction gain spans from 0 to 1, where a higher prediction gain indicates that the explainer shows a stronger shift towards a target class t.

The *Realism* is a metric used to quantify how closely a data instance fits a known data distribution. Since in our work, we employ DP to avoid providing CF that may expose statistical information about the training set, we employ realism to quantify how well CFs and private CFs fit in with the known data distribution. Furthermore, realism is employed to compare a data point with its associated CF, to assess the authenticity of CFs and to examine the impact of the data perturbation with DP on CF. Inspired by the strategies proposed in [25, 37], we train a denoising autoencoder on the noised training set, and compute the realism of a data point as the reconstruction error as the mean squared error (MSE) of the autoencoder (Eq. 3). A lower value of $realism$ corresponds to a better fit of the data point within the data distribution.

$$\text{Realism} = \frac{1}{N} \sum_{i=1}^{N} \|\text{input}_i - \text{reconstruction}_i\|^2 \tag{3}$$

Actionability is a commonly used metric that allows to measure the quality of a CF. *Actionability* assumes both the number of altered features and the degree of those changes (amount of modification in a feature) in a CF relative to the original data instance. We compute *Actionability* by taking the L1 norm of the absolute difference between the data point and its corresponding CF (Eq. 4. A low value of actionability (desired) indicates that a relatively small subset of the input features has been perturbed, with respect to the original data instance, to achieve a different outcome by the classifier, and hence, a low *actionability* suggests better quality CFs.

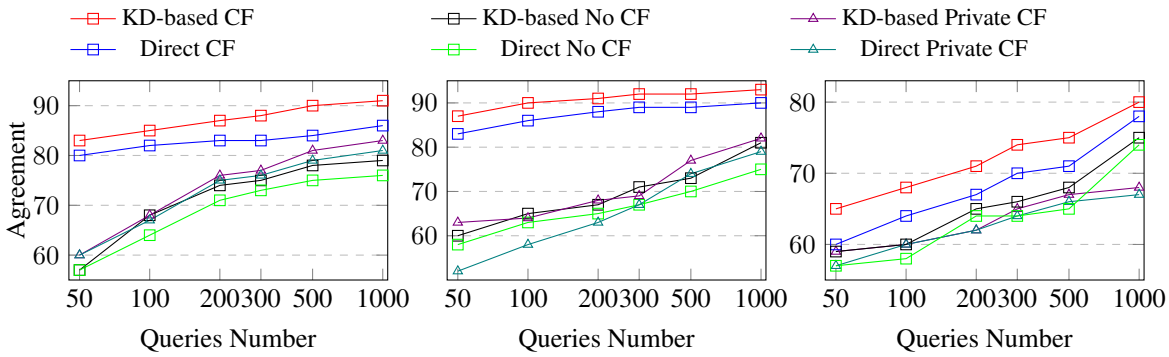
$$\text{Actionability} = \frac{1}{N} \sum_{i=1}^{N} \|\text{input}_i - \text{CF}_i\|^1 \tag{4}$$

By computing this comprehensive set of metrics, we aim to explore how DP impacts the feasibility of taking actionable steps and how this influence contributes to the prediction gain outcomes.

6 Results and evaluation

We start our discussion by analyzing the *agreement* obtained by the employed approaches for MEA in the various scenarios. During the training phase of the threat model, the attacker interacts with the model by providing it with randomly generated data points. In our experimentation, we systematically increase the number of queries to conduct a comprehensive analysis of their impact on MEA.

Figure 3 shows the *agreement* achieved by *KD-based MEA* and *Direct* in the various scenarios with respect to the number of queries made to the API across the GMSC, the Credit Card Fraud, and the California Housing datasets. Results show a consistent pattern across all scenarios as agreement initially increases significantly as the number of queries increases until a point where additional queries cease to yield a notable increase in agreement. This suggests that there is a saturation point in the MEA, and further queries may not necessarily contribute to enhancing the agreement.



(a) GMSC Dataset (b) Credit Card Fraud Dataset (c) Housing Dataset

Figure 3: The agreement values achieved by the MEA approach in the various scenarios across the a) GMSC dataset, b) Credit Fraud Dataset, c) Housing Dataset with respect to number of queries made.

We now focus on analyzing the effectiveness of exploiting CFs for performing MEAs, comparing the agreement achieved by each of *KD-based MEA* and *Direct* when CF explanations are used to that when only data points are used (i.e., comparing *Direct No CF* to *Direct-CF*, and *KD-Based No CF* to *KD-based CF*). In both cases, when CFs are exploited, the MEA is more effective, achieving an agreement significantly higher than its counterpart. For instance, in GMSC dataset, *Direct-CF* achieves an agreement ranging between 80 and 85 while *Direct No CF* achieves an agreement ranging between 55 and 75 (showing an additional agreement of up to 25%). For Credit Card Fraud dataset, the difference is significantly larger as *Direct-CF* shows an agreement ranging between 83 and 90 while Direct-no-CF has an agreement ranging between 58 and 75. This difference is also observed for California Housing dataset where exploiting CFs allows for an additional agreement of around 10%. Similarly, analyzing the agreement achieved when employing our proposed approach with CFs to its counterpart (i.e., *KD-based CF* vs. *KD-Based No CF*) across the three data sets, we notice that exploiting CFs permits obtaining up to additional 27% of agreement (for GMSC dataset at 50 queries) and, in the worst case, an additional 4% on agreement (for Housing dataset at 1000 queries). These results confirm previous findings in literature (e.g., [12] and [60]) which show the impact of using model's CF explanations for performing MEA.

We now compare our proposed approach (*KD-based MEA*) to *Direct* in each of the three scenarios (no CF, CF, and Private CF). Results show that in all scenarios, and across all data sets, our proposed approach outperforms *Direct*. In the case of *KD- based No CF* maintains an additional 3% to 6% of agreement across all data sets, with respect to *Direct no CF*. In terms of number of queries, *KD-based CF* achieves an agreement of 82% and 85% with only 50 and 100 queries, a performance that is only reached by *Direct CF* with x6 number of queries (300) and x10 number of queries (500), respectively. A similar finding can be extracted from the Credit Card Fraud dataset, as KD-CF achieves with 100 queries an agreement of 90%, which is only attainable by Direct-CF using 1000 queries. Such findings are also attainable by analyzing the performance achieved by these two scenarios with the California Housing data set.

In the case when CFs are provided by the MLaaS and are then exploited by the attacker, i.e., *KD-based CF* vs. *Direct CF*, results show that across all the 3 datasets, the agreement achieve by *KD-based CF* agreement always surpasses the Direct-CF, independent of the number of queries. In the case *KD-based CF* maintains an additional 2% to 6% agreement across all datasets. Similarly, in the case where the MLaaS provides private CFs, i.e., comparing *KD-based Private CF* to *Direct Private CF*, *KD-based Private CF* maintains a slightly additional range from 1% to 2% in most query numbers across all datasets. The observed decrease in agreement of MEA with private CFs may be attributed to the privacy-preserving nature of DP. Differential privacy introduces intentional noise into the data to protect individual privacy, making it more challenging for attackers to accurately extract sensitive information. In the context of *KD-based MEA*, when DP is employed, the noise introduced during the training process to achieve privacy guarantees could interfere with the teacher-student model's ability to transfer knowledge much more effectively.

In this consideration, using CFs significantly reduces the number of instances required for MEA to reach a specific high agreement level. For example, on the GMSC dataset, achieving 80% agreement with CFs took just 50 queries, while accomplishing the same level without CFs required over 1000 queries. This pattern holds across datasets, such as the Credit Card Fraud Dataset, where 83% agreement needed only 50 CF queries compared to over 1000 queries without CFs. In essence, the use of CFs consistently leads to a substantial reduction in the number of queries needed to achieve high agreement levels across diverse datasets. In addition to that, our proposed methodology KD-based MEA consistently results in higher agreement levels of MEA compared to the baseline approach. This holds in various scenarios and is independent of whether the MLaaS provides users with CF explanations, and across diverse datasets and query numbers. This analysis emphasizes the effectiveness of our proposed method, utilizing KD to achieve superior agreement levels. Importantly, when CFs are employed, MEA requires significantly fewer instances to reach a specific high agreement level. This highlights the efficiency of incorporating CFs in conjunction with KD to enhance the performance of the MEA.

We now analyze the impact of generating private CFs through incorporating DP in the explainer, i.e., by incorporating DP during the generation process of explanations, on mitigating MEA. Specifically, we aim to quantify the extent of protection offered by providing users with private CFs (as opposed to non-private CFs) such as to prevent attackers from leveraging vulnerabilities associated with non-private CFs. To this end, we compare the performance of each of *KD-based CF* and *Direct CF* to its private CF counterparts (*KD-based Private CF* and *Direct Private CF*, respectively). For the GMSC dataset, results show that for the relatively low number of queries (50 to 200), the difference is vast between each scenario with private CF and its counterpart. For instance, *KD-based Private CF* achieves an agreement ranging between 60% and 75% while that *KD-based CF* obtains an agreement between 83% and 87%.

In the case of the GMSC dataset, we observe that the incorporation of DP shows that our proposed strategy can maintain agreement levels comparable to scenarios without CFs and less than when incorporating traditional CFs, with an agreement range of 60% to 83% with *KD-based CF* and 60% to 81% with direct training, compared to 57% to 79% with *KD-based MEA no CF* and a range of 57% to 76% with *Direct No CF*. For the credit card fraud dataset, a similar trend is observed. With *KD-based Private CF*, we achieve a range of 63% to 82% and 52% to 79% for a direct-DP train, compared to *KD-based no CF* of 60% to 81% and a range of 58% to 75% with *Direct No CF*. For the California housing dataset, the agreement of *KD-based CF* ranges from 59% to 68% with KD and 57% to 67% with *Direct Private CF*, compared to 59% to 75% with *KD-based no CF*, and direct no CF ranges from 57% to 74%. It is also worth noting that, *KD-based Private CF*is slightly better than direct training over all datasets.

Results suggest that the incorporation of DP can play a crucial role in maintaining agreement levels comparable to scenarios without CFs, as they show that across the three datasets a mitigation can be achieved. In the appendix, we add further discussion about the impact of DP on the generation of CF and discuss what could be the potential reasons behind such results.

7 Conclusion

In this paper, our primary focus is on elucidating the privacy implications associated with providing Counterfactual Explanations (CFs) within a Machine Learning as a Service (MLaaS) platform. We first examine how CFs can be exploited for performing effective Model Extraction Attacks (MEA). We propose a novel approach based on Knowledge Distillation (KD) to effectively train a substitute model. Then, we propose to incorporate Differential Privacy (DP) into the CF generator and we assess to what extent it can serve as a mitigation strategy against MEAs, and explore its potential influence on the quality of generated CFs. Experimental results affirm that CFs can be vulnerable to exploitation for MEA. Moreover, results show that our proposed KD-based MEA approach outperforms the baseline approach. The results also demonstrate that integrating DP into the CF generation process effectively mitigates MEA while preserving to an acceptable extent the quality of CF explanations. This work represents a crucial step toward the dual objectives of preserving privacy and maintaining the integrity of explanation quality. For future work, we aim to conduct a more in-depth analysis of the impact of DP on the CF generation process, and its role in mitigating other types of attacks. This will involve a sensitivity analysis, focusing on the impact of noise levels and the guarantees provided by DP in the context of CF generation.

References

[1] Aws explainable ai. https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html. Accessed: 2024-01-04.

[2] California housing dataset. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html. Accessed: 2024-01-04.

[3] Credit card fraud dataset. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Accessed: 2024-01-04.

[4] Give me some credit dataset. https://www.kaggle.com/c/GiveMeSomeCredit. Accessed: 2024-01-04.

[5] Give me some credit notebook. https://www.kaggle.com/code/bannourchaker/credit-deep-learning. Accessed: 2024-01-04.

[6] Google auto ml. https://cloud.google.com/automl?hl=en. Accessed: 2024-01-04.

[7] Google explainable ai. https://cloud.google.com/explainable-ai. Accessed: 2024-01-04.

[8] Ibm explainable ai. https://aix360.mybluemix.net/. Accessed: 2024-01-04.

[9] Microsoft azure ai. https://azure.microsoft.com/en-us/products/machine-learning. Accessed: 2024-01-04.

[10] Microsoft explainable ai. https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability?view=azureml-api-2&WT.mc_id=docs-article-lazzeri. Accessed: 2024-01-04.

[11] 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*, pages 308–318, 2016.

[12] U. Aïvodji, A. Bolot, and S. Gambs. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

[13] Z. Chen, F. Silvestri, G. Tolomei, H. Zhu, J. Wang, and H. Ahn. Relace: reinforcement learning agent for counterfactual explanations of arbitrary predictive models. *arXiv preprint arXiv:2110.11960*, 2021.

[14] J. H. Cho and B. Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[15] F. K. Došilović, M. Brčić, and N. Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[16] C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[17] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[18] F. Ezzeddine, O. Ayoub, D. Andreoletti, and S. Giordano. Sac-fact: Soft actor-critic reinforcement learning for counterfactual explanations. In *World Conference on Explainable Artificial Intelligence*, pages 195–216. Springer, 2023.

[19] M. Hashemi and A. Fathi. Permuteattack: Counterfactual explanation of machine learning credit scorecards. *arXiv preprint arXiv:2008.10138*, 2020.

[20] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[21] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[22] C. Huang, M. Pawelczyk, and H. Lakkaraju. Explaining the model, protecting your data: Revealing and mitigating the data privacy risks of post-hoc model explanations via membership inference. In *ICML 2024 Next Generation of AI Safety Workshop*.

[23] C. Huang, C. Swoopes, C. Xiao, J. Ma, and H. Lakkaraju. Accurate, explainable, and private models: Providing recourse while minimizing training data leakage. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.

[24] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USENIX Security 20)*, pages 1345–1362, 2020.

[25] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[26] M. Juuti, S. Szyller, S. Marchal, and N. Asokan. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 512–527. IEEE, 2019.

[27] K. Kanamori, T. Takagi, K. Kobayashi, Y. Ike, K. Uemura, and H. Arimura. Ordered counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11564–11574, 2021.

[28] K. Krishna, G. S. Tomar, A. P. Parikh, N. Papernot, and M. Iyyer. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations*, 2020.

[29] S. Krishna, J. Ma, and H. Lakkaraju. Towards bridging the gaps between the right to explanation and the right to be forgotten. In *International Conference on Machine Learning*, pages 17808–17826. PMLR, 2023.

[30] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[31] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS quarterly*, 38(1):73–100, 2014.

[32] S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

[33] T. Miura, T. Shibahara, and N. Yanai. *MEGEX: Data-Free Model Extraction Attack Against Gradient-Based Explainable AI*, page 56–66. Association for Computing Machinery, New York, NY, USA, 2024.

[34] J. Moore, N. Hammerla, and C. Watkins. Explaining deep learning models with constrained adversarial examples. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I 16*, pages 43–56. Springer, 2019.

[35] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[36] 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)*, pages 739–753. IEEE, 2019.

[37] D. Nemirovsky, N. Thiebaut, Y. Xu, and A. Gupta. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In *Uncertainty in Artificial Intelligence*, pages 1488–1497. PMLR, 2022.

[38] D. Nguyen, N. Bui, and V. A. Nguyen. Feasible recourse plan via diverse interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 4679–4698. PMLR, 2023.

[39] T. M. Nguyen, T. P. Quinn, T. Nguyen, and T. Tran. Counterfactual explanation with multi-agent reinforcement learning for drug target prediction. *arXiv preprint arXiv:2103.12983*, 2021.

[40] T. T. Nguyen, T. T. Huynh, Z. Ren, T. T. Nguyen, P. L. Nguyen, H. Yin, and Q. V. H. Nguyen. A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures. *arXiv preprint arXiv:2404.00673*, 2024.

[41] D. Numeroso, D. Bacciu, et al. Explaining deep graph networks with molecular counterfactuals. In *Proceedings of the NeurIPS2020 Workshop on Machine Learning for Molecules*, 2020.

[42] S. J. Oh, B. Schiele, and M. Fritz. Towards reverse-engineering black-box neural networks. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144, 2019.

[43] A. C. Oksuz, A. Halimi, and E. Ayday. Autolycus: Exploiting explainable ai (xai) for model extraction attacks against decision tree models. Network and Distributed System Security Symposium (NDSS), 2023.

[44] T. Orekondy, B. Schiele, and M. Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4954–4963, 2019.

[45] S. Pal, Y. Gupta, A. Shukla, A. Kanade, S. Shevade, and V. Ganapathy. Activethief: Model extraction using active learning and unannotated public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 865–872, 2020.

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

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

[48] S. Pentyala, S. Sharma, S. Kariyappa, F. Lecue, and D. Magazzeni. Privacy-preserving algorithmic recourse. *arXiv preprint arXiv:2311.14137*, 2023.

[49] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[50] 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*, pages 1135–1144, 2016.

[51] S. Sharma, J. Henderson, and J. Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[52] P. P. Shinde and S. Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

[53] 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*, pages 231–241, 2021.

[54] 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)*, pages 3–18. IEEE, 2017.

[55] K. Sokol and P. Flach. Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety. In *2019 AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019*. CEUR Workshop Proceedings, 2019.

[56] D. C. Thang, N. T. Tam, N. Q. V. Hung, and K. Aberer. An evaluation of diversification techniques. In *Database and Expert Systems Applications: 26th International Conference, DEXA 2015, Valencia, Spain, September 1-4, 2015, Proceedings, Part II 8*, pages 215–231. Springer, 2015.

[57] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016.

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

[59] K.-C. Wang, Y. Fu, K. Li, A. Khisti, R. Zemel, and A. Makhzani. Variational model inversion attacks. *Advances in Neural Information Processing Systems*, 34:9706–9719, 2021.

[60] 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*, pages 1318–1329, 2022.

[61] A. Yan, R. Hou, X. Liu, H. Yan, T. Huang, and X. Wang. Towards explainable model extraction attacks. *International Journal of Intelligent Systems*, 37(11):9936–9956, 2022.

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

[63] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 5(11):1364–1375, 2012.

A Appendix / supplemental material

A.1 Preliminaries

In this section, we discuss the key concepts used in our work. We first introduce the concept of CF explanations. Then, we provide an overview of DP. Finally, we introduce the concept of KD.

A.1.1 Counterfactual Explanations (CF)

A CF explanation is a type of example-based explanation that provides a hypothetical scenario that illustrates how a different decision or outcome could have arisen if the input data had been altered [58]. CFs provide users with an actionable explanation as they enable users to understand how changes in inputs would affect the model's output [49]. Additionally, CFs allow to identification of the variables that should have differed in a given input to observe a different outcome, thus making it possible to assess the influence of specific factors, which in turn provides valuable insights into how decisions are made by an ML model [58]. This type of CF analysis has already proven to improve the interpretability of ML models in several domains, such as healthcare and finance [15].

To generate CFs, various CF explanation methods obtain CFs by optimizing a customized cost function. These explainers employ distinct strategies such as mixed-integer linear optimization [27], heuristic search strategies [31, 34], and metaheuristic approaches [19, 39, 41, 51], such as genetic algorithms [19, 51], Reinforcement Learning [13, 18], Generative Adversarial networks [37] or Graph density bases [49]. In our work, we employ the method of Generating CFs for Real-Time Recourse and Interpretability using Residual GANs (CounteRGAN) as a CF explainer [37]. The authors of CounteRGAN formalized a novel residual generative adversarial network that trains the generator to produce residuals that are intuitive to the notion of perturbations used in counterfactual searches. The search process seeks to maximize the value function concerning the discriminator D and minimize it concerning the generator G. Where C_t is the target classifier to be explained, Reg(G, x_i) is a regularization term, and x_i are samples drawn from the entire data distribution [37].

$$
\begin{aligned}
V_{\text{CounterGAN-bb}}(D, G) = & \frac{\sum_i C_t(x_i) \log D(x_i)}{\sum_i C_t(x_i)} \\
& + \frac{1}{N} \sum_i \log(1 - D(x_i + G(x_i))) \\
& + \text{Reg}(G, \{x_i\}),
\end{aligned}
\tag{5}
$$

CF generators aim to find in-distribution points by optimizing metrics of actionability such as sparsity and proximity [35]. Sparsity measures how many features of the CF data point are different from the original data point. Proximity measures the overall distance between the CF and the original data point, where a low proximity value indicates that the two data points are similar. This can be useful for providing actionable feedback, as it suggests that the model is recommending only a few changes that are likely to have a significant impact on the output.

A.1.2 Knowledge Distillation

Knowledge distillation, also known as model distillation, is a process that involves the transfer of knowledge from a high-complex model, known as a *teacher model*, to a reduced-complexity model, known as a *student model*, that remains operable within real-world constraints [14]. While KD constitutes a specific instance of model compression, it serves as a means to extract essential insights, patterns, and expertise embedded in the larger models to create a deployable model. The student model follows a training procedure with the primary objective of imitating the performance of the teacher model. This knowledge transfer mechanism involves the student model learning not only the surface-level predictions made by the teacher model but also the deeper patterns, generalizations, and decision-making strategies embedded within the teacher architecture. To achieve this, the student model is guided during training by incorporating two primary sources of information: the actual target labels or predictions for the dataset at hand, and the soft labels or probability distributions generated by the teacher model in response to the same dataset.

KD necessitates the availability of a well-trained teacher model, a trainable student model, and the specification of a student loss function for assessing predictions against ground-truth labels (L). A distillation loss function, accompanied by a temperature parameter ($temp$), is employed to bridge the knowledge gap by comparing the soft predictions of the student to the softened teacher labels. The student and the distillation losses are weighted via an alpha (α) which is essential for balancing task-specific accuracy and knowledge transfer. Subsequently, the losses incurred are computed, incorporating a weighted combination of the student loss (weighted by α) and the distillation loss (weighted by $1 - \alpha$). This weighting mechanism allows for a fine-tuned balance between preserving task-specific accuracy (student loss) and incorporating the knowledge distilled from the teacher model (distillation loss). The Kullback-Leibler (KL) divergence, denoted as $KL(P \parallel Q)$ (Eq. 6), is a mathematical measure employed in KD to assess the difference between two probability distributions, P and Q, where the summation is taken over all classes or categories.

Suppose P represents the soft probabilistic predictions of a teacher model, while Q represents the corresponding probabilistic predictions of a student model (expressed as the output made by a softmax activation function). During training, KL divergence as the distillation loss, encourages the student model to capture the nuances and uncertainties present in the teacher's predictions. In summary, the loss to be optimized by the student is shown by Eq. 7.

$$distillation\_loss = KL(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \tag{6}$$

$$loss = \alpha \cdot student\_loss + (1 - \alpha) \cdot distillation\_loss \tag{7}$$

A.1.3 Differential Privacy

Differential privacy is a mathematical framework applied to safeguard individual records by introducing controlled noise to the data and allowing the extraction of valuable insights while ensuring that individual identities remain protected [17]. This enables analyzing data without disclosing sensitive information about any specific individual in the dataset [16]. DP aims to guarantee, by definition, that the inclusion or exclusion of any individual record in the dataset should have minimal impact on the outcome of the mechanism. A mechanism refers to any mathematical computation applicable to and with the data. In the context of classification tasks, DP analyzes how the output undergoes probabilistic changes based on a given input. Hence, a mechanism guarantees differential privacy if the likelihood of any outcome is nearly identical for any two datasets that vary by only one record.

One widely utilized approach for addressing numerical inquiries involves the incorporation of randomized approaches that add calibrated random noise, which works by introducing sufficient noise to the input or output of the mechanism to obscure the potential contribution of any individual record in the data while simultaneously maintaining the overall accuracy of the analysis. In Def. A.1, we detail the DP inequality.

Definition A.1 (Differential Privacy) *A randomized algorithm M with domain $\mathbb{N}^2$ is (ε, δ)-differentially private if for all $S \subseteq Range(M)$ and for all $x, y \in \mathbb{N}^2$ such that $||x - y||_1 \leq 1$:*

$$\Pr[M(x) = S] \leq e^{\varepsilon} \cdot \Pr[M(y) \in S] + \delta,$$

Def. A.1 states that an algorithm M (ε, δ) is differentially private if, for all subsets S of the range of M and all pairs of inputs x and y differing by at most one data instance, the probability of $M(x)$ outputting S is bounded by a factor of e^{ε} and an error margin δ.

In addition to the perturbation of input or output data, it is possible to achieve DP during the training of ML models and safeguard the privacy of the training model itself. This is usually achieved by introducing perturbations to the model weights and gradients during training [11, 63]. The process proceeds as follows: consider an ML model, to be trained on dataset D, with a parameter set w^* will minimize an objective $L_D(w) = \sum_{t_i \in D} L(t_i, w)$, DP is injected during the optimization process to prevent the model from memorizing specific details about individual data points. DP training involves multiple steps of adding noise to the gradient of the model parameters concerning the training data being trained on, followed by gradient clipping. Then, the resulting parameter set w^*, minimizes the loss that can be later derived.

Table 1: Actionability and Prediction Gain of CFs generated on the random set with and without the adoption of DP in the generator.

	Prediction Gain		Actionability	
Data	CFs	Private CFs	CFs	Private CFs
GMSC	0.243 ± 0.011	0.121 ± 0.01	24.567 ± 0.364	16.981 ± 0.158
Credit Fraud	0.700 ± 0.084	0.445 ± 0.06	35.269 ± 0.328	10.507 ± 0.238
Housing	0.633 ± 0.052	0.678 ± 0.024	3.852 ± 0.053	1.004 ± 0.016

A.2 Datasets

We perform our experiments using three classification datasets, namely, Give Me Some Credits dataset [4], Credit Card Fraud Dataset [3], and California Housing Dataset [2]. The datasets are described as follows.

- **Give Me Some Credits [4]** This dataset is collected to forecast the likelihood of an individual undergoing financial distress within the next two years based on their financial and demographic features. The full dataset encompasses 150,000 applicants, with 139,974 applicants classified as good and 10,026 applicants labeled as bad. We follow the pre-processing of [5].
- **Credit Card Fraud [3]** The dataset contains transactions conducted by European cardholders using credit cards in September 2013. Within this dataset, transactions recorded over two days reveal the occurrence of 492 fraudulent instances out of a total of 284,807 transactions.
- **California Housing [2]** A dataset was created based on the 1990 U.S. census, consisting of 8 features and a target variable representing the median house value for California districts in dollars. The target variable is then converted into two classes using a threshold defined by the median.

A.3 Impact of Incorporating DP in CF generator on quality of explanations

We shift our focus to analyzing the impact of incorporating DP within the CounteRGAN CF generator on the quality of CFs[4] in terms of the metrics introduced in Sec. 5.5.

Table 1 reports the prediction gain of CounteRGAN and the actionability of the CFs achieved with and without DP across the three datasets. Note that the focus lies on comparing the metrics on the CFs exclusively between scenarios with and without DP, regardless of the applied MEA approach. Results show that CounteRGAN achieves a higher prediction gain (a higher probability shift, desired) when DP is not incorporated within the explainer than when employed for GMSC (0.243 ± 0.011 versus 0.121 ± 0.01) and Credit Fraud datasets (0.700 ± 0.084 versus 0.445 ± 0.06) while for Housing dataset, CounteRGAN shows a prediction gain slightly higher in the case where DP is incorporated rather than when not (0.678 versus 0.633). These findings highlight that the integration of DP has an impact on prediction gain, constraining the explainer's progress toward the desired class.

In terms of actionability which is measured as the degree of perturbation or modification of the CF compared to the initial point, the results show that the explainer generates CFs with lower values of actionability when DP is incorporated than when not across all datasets. For instance, for GMSC dataset, the actionability of CFs with DP is 16.981 ($\pm$ 0.158) while that without DP is 24.5 ($\pm$ 0.364). Similarly, for credit fraud and housing datasets, the actionability is reduced when incorporating DP from 35.269 ± 0.328 to 10.507 ± 0.238, and from 3.852 ± 0.053 to 1.004 ± 0.016 respectively. This means that that the practical usefulness of the CFs or perturbations in guiding actionable decisions is reduced, suggesting that the privacy constraints imposed might compromise the effectiveness of CF analyses in providing actionable guidance for influencing desired outcomes.

Although a higher prediction gain and lower actionability are preferable when generating CFs, the results suggest that the CF generation with DP has taken a different trajectory than it has taken without DP, leading to a reduction in prediction gain. This shift in the CF generation approach is

[4]Note that this analysis is related to the explainer and the quality of the generated CFs, not to the methodology adopted to perform MEA.

Table 2: Realism of the random points and the CFs generated with and without incorporating DP in the generator across the datasets.

	Realism		
Data	random points	CFs	Private CFs
GMSC	15.649 ± 0.033	8.56 ± 0.142	15.723 ± 0.033
Credit Fraud	3.104 ± 0.019	3.072 ± 0.1647	4.73 ± 0.027
Housing	2.070 ± 0.04	1.356 ± 0.031	2.0 ± 0.01

reflected in the probability, ultimately contributing to the observed decrease in agreement, thereby motivating the smaller agreement for MEA.

To investigate the impact of incorporating DP in the explainer in more detail, we analyze the realism (which allows us to quantify how well a data instance fits a data distribution of a dataset) of the data points, including CFs in the scenarios (no CF, CF, Private CF).

Table 2 reports the realism of the data points used as initial queries and the CFs generated by the explainer across the three data sets. Results show that the Realism of CFs without incorporating DP in the generator is the lowest in comparison to the Realism of random points and that of CFs when DP is employed, meaning that the CF generator has produced more realistic data points (lower Realism), attempting to align the CFs of random points with the distribution of the training data. In particular, for the GMSC dataset, the randomly generated points have an average realism of 15.6 ($\pm$ 0.03), and the CF generator has produced corresponding CFs with a realism of 8.5 ($\pm$ 0.14). Similarly, for the Credit Fraud data set, the Realism of random points is 3.104 ($\pm$ 0.019) while for the corresponding CFs, it is slightly lower at 3.07 ($\pm$ 0.16), respectively. For the Housing data set, the realism of data points is 2.07 ($\pm$ 0.04) while their corresponding CFs had a realism of 1.35 ($\pm$ 0.03). This implies that the explainer generates CFs with lower realism from the original queried random points and aligns them more closely with the distribution of the training data. The private CFs preserve a realism very similar to that of random points. More specifically, for the GMSC dataset, the private CFs have preserved their initial average level of realism of 15.723 ($\pm$ 0.033). Similarly, the housing dataset preserves a realism of 2.0 ($\pm$ 0.01). For the Credit fraud dataset, the realism of the CFs deviates by one degree from the distribution, resulting in a realism of 4.73 ($\pm$ 4.73). With the incorporation of DP, realism can be preserved indicating the effectiveness of the Private CFs in maintaining the quality and distribution of the original data. The findings might vary depending on the dataset, as shown by the slight deviation in realism for the Credit Fraud dataset.

It is worth noting that, random points are inherently unrealistic (do not exhibit high realism). Our private CF generation approach ensures this unrealistic nature is preserved when queried with a random data point. This is crucial because our experiments showed that traditional CFs, without privacy protections, generate more realistic outputs from random data, potentially revealing private information.

A.4 Limitations and Future Work

The limitations of our current approach and potential avenues for future research are the following:

- **Attack Scope**: Our work focuses on MEA leveraging CFs, it is important to acknowledge the potential existence of other privacy attacks such as membership inference and model inversion that warrant further investigation. To this end, in future work, we plan to evaluate the effectiveness of our proposed private GAN-based CF generation approach against other attack types.
- **Privacy-Performance Trade-off**: In this work, we made the first attempt to integrate DP in GAN-based CF generators. We assume that GAN-based CFs should be privacy-preserving against random and zero-knowledge queries. This means it should not reveal any realistic CFs when queried by random noise. A future direction is to thoroughly evaluate the impact of various DP levels on considering both their influence on mitigating various privacy attacks including MEA and their impact on the effectiveness and quality of CFs with DP.
- **Focus on Deep Learning Applications**: KD is effective with DNNs due to its complex architecture and ability to learn rich data representations. Future research aims to explore

adapting KD for other ML algorithms such as tree and ensemble-based algorithms and to potentially unlock similar performance enhancements.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarized our contributions in the abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations are discussed as a separate section in the appendix.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: All assumptions stated are referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the steps required for reproducibility are provided in the experimental settings.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use publicly available datasets (references to datasets are reported in the paper). We provide sufficient instructions and code to reproduce the main experimental results (libraries, parameters, etc.).

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the necessary informations.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the results of the average of 10 runs with randomly chosen subsets for each experiment of the MEA KD-based attack.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Experiments were computed with a machine of intel Core i7, a GPU of RTX 3070, and 8 GB of RAM.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the ethics guideline.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We discuss the effect of performing a model extraction attack, and we propose a mitigation method.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available datasets and codes from Kaggle that we referenced.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not provide new assets.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.