

# On the interplay of Explainability, Privacy and Predictive Performance with Explanation-assisted Model Extraction

Fatima Ezzeddine<sup>1,2,\*†</sup>, Rinad Akel<sup>3†</sup>, Ihab Sbeity<sup>3</sup>, Silvia Giordano<sup>2</sup>, Marc Langheinrich<sup>1</sup> and Omran Ayoub<sup>2</sup>

<sup>1</sup>Università della Svizzera italiana, Lugano, Switzerland

<sup>2</sup>University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland

<sup>3</sup>Lebanese University, Beirut, Lebanon

## Abstract

Machine Learning as a Service (MLaaS) has gained important attraction as a means for deploying powerful predictive models, offering ease of use that enables organizations to leverage advanced analytics without substantial investments in specialized infrastructure or expertise. However, MLaaS platforms must be safeguarded against security and privacy attacks, such as model extraction (MEA) attacks. The increasing integration of explainable AI (XAI) within MLaaS has introduced an additional privacy challenge, as attackers can exploit model explanations—particularly counterfactual explanations (CFs) to facilitate MEA. In this paper, we investigate the trade-offs among model performance, privacy, and explainability when employing Differential Privacy (DP), a promising technique for mitigating CF-facilitated MEA. We evaluate two distinct DP strategies: implemented during the classification model training and at the explainer during CF generation.

## Keywords

Counterfactual Explanations, Model Extraction Attack, Differential Privacy

## 1. Introduction

Machine Learning (ML) as a Service (MLaaS) is becoming increasingly popular for deploying powerful predictive models as it facilitates access to ML training and deployment tools, while eliminating the need for extensive computational resources [1]. The adoption of MLaaS, however, introduces important security and privacy risks. For instance, adversaries can query the deployed ML models through application programming interfaces (APIs) to perform various types of attacks, such as membership inference (MIA) [2] and model extraction (MEA) [1]. These attacks, if successful, pose serious threats to data privacy and intellectual property. For instance, MIA can reveal whether specific data points were used in training, MEA, instead, enables adversaries to replicate proprietary models, leading to financial losses and competitive disadvantages and facilitates other data privacy attacks by having access to a copy of the model. To defend against these attacks, data privacy-enhancing technologies such as Differential Privacy (DP) [3] exist. DP has shown effectiveness in defending against such attacks and is therefore widely adopted in use cases that require data and model sharing and deployments [4]. DP enables privacy-preserving training of deep neural networks (DNN) to effectively mitigate inferential attacks by adding a controlled amount of noise to either raw data or model weights and ensures that individual data points have minimal influence on the model's response, which limits the amount of sensitive information leaked when an attacker queries the model.

Recently, with the increasing demand for transparency in automated decision-making, MLaaS platforms are starting to incorporate Explainable Artificial Intelligence (XAI) [5] techniques into their workflows to provide explanations of the model's decisions<sup>1</sup> [6, 7]. These platforms now provide not only the final decisions of ML models but also explanations of the underlying processes. The increased

---

*Late-breaking work, Demos and Doctoral Consortium, collocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey*

\*Corresponding author.

†These authors contributed equally.

✉ fatima.ezzeddine@usi.ch (F. Ezzeddine)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup><https://aws.amazon.com/sagemaker/clarify/>, <https://cloud.google.com/explainable-ai>

transparency provided by XAI introduces new challenges for preserving privacy and safeguarding MLaaS platforms from adversarial threats, as model’s explanations can inadvertently reveal information about model’s decision boundaries [8]. Specifically, counterfactual explanations (CFs) [9], which aim to identify the smallest changes to input data that would alter an ML model’s prediction to a desired outcome, can reveal the factors most influential in the model’s decision-making. Recent research has indeed explored how explanations can be leveraged to enhance the effectiveness of such attacks [8, 7, 10, 6, 11]. Complementing this, DP can also be applied at the explanation level, where it masks explanations to limit their utility to adversaries while balancing interpretability and privacy [7, 10]. As DP can have impact on predictive performance and explanation quality, and can be applied on both levels, there is a growing research to highlight the importance of DP in developing mitigation strategies that specifically address risks introduced by explanations, emphasizing the need to adapt, utilize, or extend existing defense methods to the exploitation of explainability. In this work, we focus on analyzing the mitigation framework that integrates DP at the model and at the explainer, and investigate the interplay between *i) model’s accuracy*, as DP is expected to influence model’s inference capability, *ii) privacy*, as employing DP provides resilience against attacks, and *iii) explainability*, as noise added to model or explainer may impact quality of explanations [6, 12]. We aim to quantify this interplay and extract insights on the choice of where to employ DP (at model or at explainer, or at both) and the degree of noise level to be employed to balance predictive performance, explainability and privacy.

To perform the attack, we employ a recently proposed MEA technique based on Knowledge Distillation (KD) due to its proven performance and practicality [10]. In terms of mitigation strategies, we employ DP at the ML model using Differential Private-SGD (DP-SGD) and at the explainer using a DP-based Generative Adversarial Network (GAN) [10] with varying noise levels. To this end, we investigate the following research questions (RQs):

- *RQ1: To what extent does applying DP at the model or at the explainer, or both, effectively mitigates MEA facilitated by CFs?*
- *RQ2: How does noise level in DP influence the effectiveness of MEAs that leverage CFs?*
- *RQ3: In what ways does the quality of CF explanations differ when DP is applied at the model compared to the explainer?*

## 2. Related Work

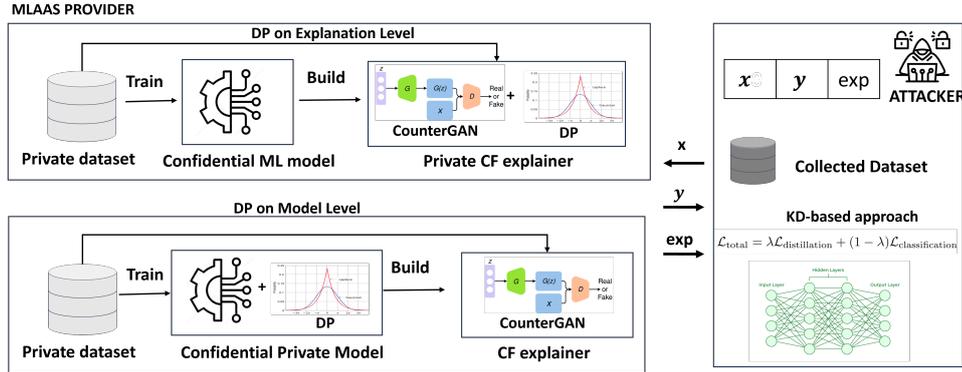
Several studies have explored leveraging XAI techniques and exploiting model explanations to perform privacy attacks. In [13], authors explore the vulnerabilities of Local Interpretable Model-agnostic Explanations and show that an adversary can generate new data samples near the decision boundary and, consequently, perform MEA by crafting adaptive queries. In [7], authors show that by leveraging gradient-based explanations, adversaries can enhance the effectiveness of MIA. In [14], the authors propose a methodology that performs MEA by jointly minimizing classification and explanation loss, thereby improving its fidelity. Other works explore the use of CFs to enhance the effectiveness of MEA. For instance, [11] introduces a methodology that relies on model predictions and CFs to train a substitute model. Similarly, [15] presents a novel strategy where CF pairs, including the CF of the CF, serve as training samples to MEA. More recently, [10] proposes a methodology based on KD techniques that exploit CFs to perform MEA effectively while minimizing the number of queries to an MLaaS system and to generate private CFs with DP. Moreover, [16] explores the theoretical foundation of MEA with CFs highlighting the risks associated with providing CF explanations.

Several approaches have been proposed to prevent adversaries from exploiting model explanations for privacy attacks. In [17], the authors propose an approach that builds on the concept of providing CFs that are not derived from the entire feature space but instead are generated within a designated space. Some works developed methodologies to generate explanations while limiting the exposure of sensitive insights related to decision boundaries, training data, or model architectures. Authors in [18] present an approach to generate differentially private CFs using functional mechanisms to protect the underlying model from potential inference attacks. In contrast, [19] proposes a novel approach that

constructs private recourse paths as CFs using differentially private clustering. Authors in [10] focus on GAN-based CF (proposed in [20]), injecting DP into the training process of the generator that is responsible for generating CFs that limits the memorization of the private data points.

Similar to these works, we focus on identifying a mitigation strategy against attacks that exploit the model’s explanations. Specifically, we explore the application of DP to the ML model, the explainer, and both simultaneously. Despite the numerous studies utilizing DP for mitigation strategies, our work is, to the best of our knowledge, the first to explore the application of DP in both the ML model and the explainer and to investigate their effectiveness in countering MEA and examine their influence on the quality of explanations. Additionally, our work explores the interplay between preserving model privacy and generating privacy-preserving CFs, as well as the implications for defending against MEA.

### 3. Problem Formulation and Methodology



**Figure 1:** Model Extraction Attack within an MLaaS provider, depicting two different scenarios where DP is employed at the model or at the explainer to counter potential attacks.

Given a dataset  $D = \{(x_i, y_i)\}_{i=1}^N$ , where  $x_i$  are feature vectors and  $y_i$  are corresponding labels. A target model  $f(x; \theta)$  trained and optimized to achieve high performance on  $D$  is deployed as MLaaS and is queryable through an API (as shown in Fig. 1). An *Attacker* (adversary) attempts to extract an approximation of  $f(x; \theta)$  using queries and the provided CFs. The attacker conducts MEA by exploiting CFs and varying the number of queries. To perform our analysis, we proceed as follows:

- Step 1: Train target models as baseline models  $f_{baseline}(x; \theta_{base})$ .
- Step 2: Generate CFs by training a CounterGAN to generate CFs  $\hat{x} = G(x; \phi)$  for  $f_{baseline}$ .
- Step 3: Simulate MEA, where the adversary queries the models with random points and collects pairs of predictions and CFs. The adversary trains an extracted model using the KD-based method proposed in [10].
- Step 4: Measure MEA success by computing the agreement on a separate dataset to quantify and compare the level of agreement between extracted and original models/explanations.
- Step 5: Assess the quality of CFs using metrics such as prediction gain, realism (explained in more details in Sec 4).

The effectiveness of the MEA is measured using similarity metrics such as agreement. In practice, this agreement expectation is estimated empirically using a set of  $n$  test inputs  $\{x_1, x_2, \dots, x_n\}$ .  $Agreement = \frac{1}{n} \sum_{i=1}^n I(f_{\theta}(x_i) = \hat{f}_{\hat{\theta}}(x_i))$ . Where  $I$  counts the number of times the extracted model’s predictions match the target model’s predictions.

As a mitigation against MEA, we employ two strategies: 1) *DP-Model with DP-SGD*: where we apply DP-SGD during model training on  $f(x; \theta)$ . 2) *DP-Explainer (DP in CounterGAN)*: We inject DP noise at the generator  $G(x; \phi)$  that outputs private CFs ([10]). We then perform MEA leveraging CFs under different DP settings, i.e., the approach adopted and the privacy parameter’s noise level  $\epsilon$ , and evaluate the adversary’s MEA success and CF quality. Specifically in step 1,  $f_{baseline}(x; \theta_{base})$  is first trained on  $D$

without DP. We also train a DP-protected model  $f_{\text{DP}}(x; \theta_{\text{DP}})$  using DP-SGD with privacy parameter  $\epsilon$ . Similarly, in step 2, we also train a private CounterGAN  $\hat{x} = G_{\text{private}}(x; \phi)$  to generate the private CFs by varying the noise level. The attacker in step 3, apply MEA to extract  $f_{\mathcal{A}}(x; \theta_{\mathcal{A}})$  using the KD-based method using either CFs generated by  $G(x; \phi)$  or  $G_{\text{private}}(x; \phi)$ . For the comparative analysis, we consider four distinct scenarios: (1) *No DP*: Baseline scenario that does not incorporate DP at any level, allowing the evaluation of the unprotected model performance and vulnerability. (2) *DP-Model*: Only the target model employs DP. This protects the model from adversarial replication while the explanation generator remains unprotected. (3) *DP-Explainer*: DP is applied to the explanation generator. This scenario assesses the impact of DP explanations on their utility without directly affecting the target model. (4) *DP-Model-Explainer*: Both the target model and the explanation generator are protected with DP, aiming to balance model performance, explanation quality, and resistance to MEA.

## 4. Experimental Settings

### 4.1. Datasets, Target and Threat Model

We perform an evaluation on 2 datasets: *Housing* [21] and *EEG Eye State*[22]. The Housing dataset describes housing prices and includes 20,640 instances and 8 features a mix of socio-economic, demographic, and geographic attributes. The target variable represents the median house value and is converted into two classes using a threshold defined by the median. The EEG Eye State dataset comprises EEG measurement data recorded using a Neuroheadset, and contains 14,980 data points and 14 features. The target variable is a binary label representing the eye-closed or opened state.

The target model  $f_{\theta}$  is a DNN with 16 hidden layers of 64, 32, 16, 32, 64, 128, 64, 32, 128, 64, 128, 64, 128, 64, 32, and 16 neurons per layer with a GELU activation function and a the softmax activation function in the output layer. We employ Adam optimizer for the cases where DP is not used and TensorFlow Privacy’s DPKerasAdamOptimizer for the cases where DP applied. The model is trained without DP and with noise levels of 0.1, 0.5 and 0.9 for DP cases and with varying learning rates (0.001, 0.002, and 0.01), and we vary the `l2_norm_clip` to between 1, and 1.5 (`l2_norm_clip` bounds the sensitivity of the gradients by limiting the influence of any single training example on the overall gradient update, which is a crucial step before adding noise). Note that the more noise, the higher the privacy. The target models are trained using 80% of the corresponding dataset, and the best-performing model in term of accuracy was chosen.

To simulate a realistic attack scenario, we assume that the attacker has no prior knowledge of the training data distribution and does not know the architecture of the target model, but can build a simple threat model  $t_{\gamma}$ . The  $t_{\gamma}$  consists of 5 layers, with 32, 64, 128, and 64 neurons with ReLU activation, followed by a softmax output layer. Attacker generate random different data points to query the model, within a range of -3 to 3 for each feature and extract CFs to feed as input to the KD-based MEA. Our evaluation involves performing MEA while varying the number of queries from 50 to 1000 and therefore the input to KD. For optimization, we utilize both Adam for the cases where DP is not used and TensorFlow Privacy’s DPKerasAdamOptimizer for the cases where DP is used to assess model performance under three different noise levels, 0.1, 0.5, and 0.9. We tune different KD-based approach hyperparameters, specifically, alpha within the range of 0.1 and 0.5, temperature within the range of 1 and 10. We compute the MEA agreement over the 20% test set and report average results of 5 runs.

### 4.2. Counterfactual generator

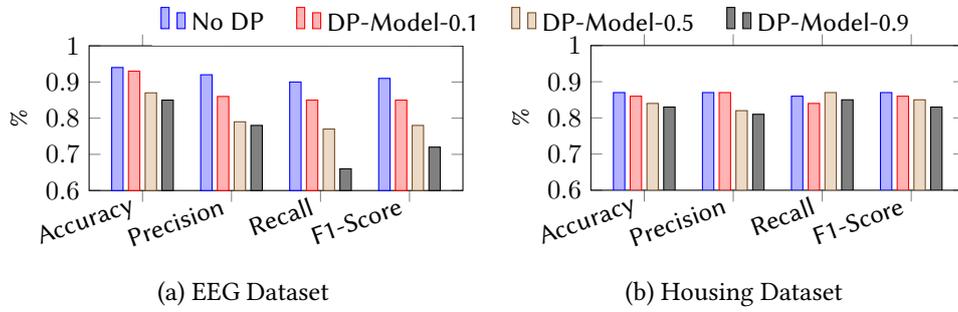
The generator of CounterGAN takes an input feature vector and processes it through 4 layers with 64, 32, and 64 neurons, with ReLU as an activation function and a final layer with Tanh activation. The discriminator follows a simple feedforward design, consisting of 128, 128, 64 neurons with ReLU activation, and the final output layer with Sigmoid activation. In the No-DP scenario, we used the standard Adam optimizer. For the scenarios where DP is employed, we applied DP using noise levels of 0.1, 0.5, and 0.9, respectively on the generator, with TensorFlow Privacy’s DPKerasAdamOptimizer. We

varied the learning rate where we used 0.05, 0.005, 0.01, and 0.001, and `l2_norm_clip` to between 1, 1.5, and 3. We report the results of the average of 5 runs. We consider the following metrics to assess the influence of employing privacy on the CFs.

- **Prediction Gain:** quantifies how the explainer modifies the input to influence the model’s decision by measuring the change in the classifier’s confidence score for a specific target class  $t$  when replacing the original data point with its CFs:  $\Delta P = P_f(CF, t) - P_f(X, t)$  Where:  $P_f(CF, t)$  is the probability score for the target class  $t$  of the CF and  $P_f(X, t)$  is for initial point.
- **Realism:** quantifies how a data instance fits within a data distribution to evaluate how well CFs and private CFs with different noise applied match the original training data distribution. It is defined as:  $Realism = \frac{1}{N} \sum_{i=1}^N \|\text{input}_i - \text{reconstruction}_i\|^2$  Where:  $\text{input}_i$  represents the original data point,  $\text{reconstruction}_i$  is the corresponding autoencoder reconstruction and  $N$  is the total number of instances ([20]). A lower realism value indicates that the data point is more realistic.

## 5. Results and Discussion

### 5.1. ML Model Predictive Performance

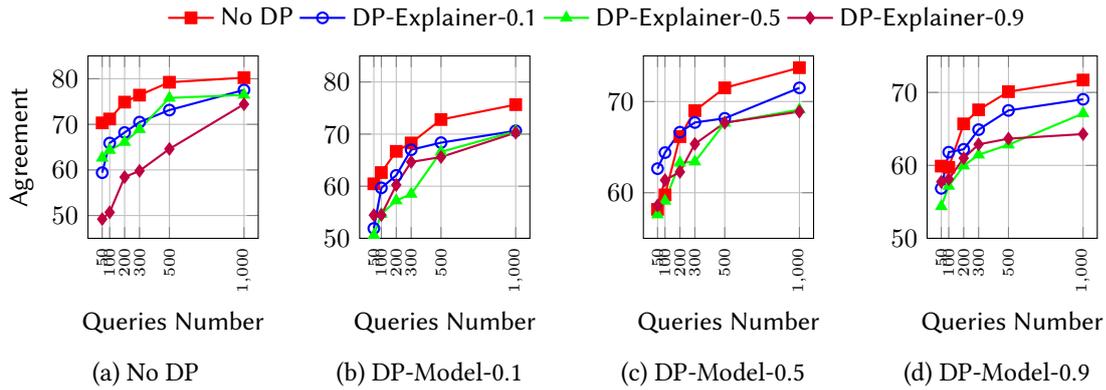


**Figure 2:** Model Performance achieved by the ML model across the two datasets for varying noise scales.

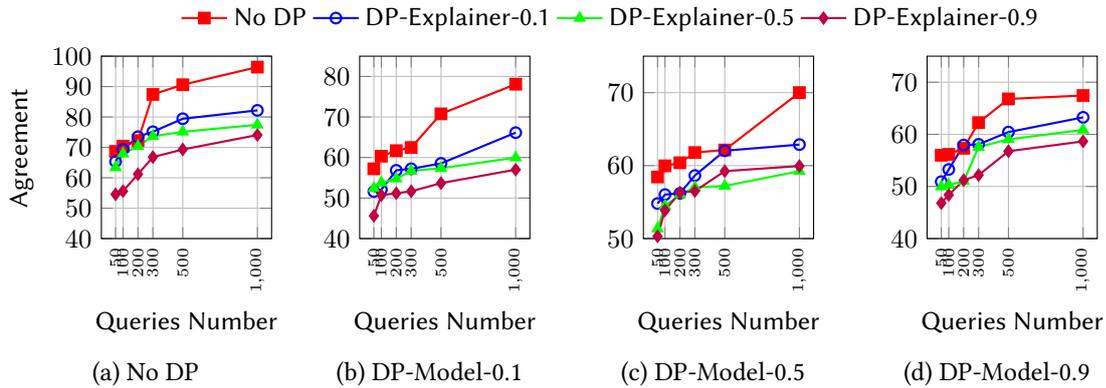
Fig 2 reports the predictive performance metrics of the models while varying the noise level across the two datasets used in our evaluations. As previously mentioned, we consider three noise levels when applying DP, 0.1, 0.5 and 0.9, and we refer to each case as DP-Model-noise level. As expected, the results across the two datasets indicate a decline in predictive performance metrics as the noise level increases. For instance, in the EEG dataset, accuracy, precision, recall, and F1-score are 0.94, 0.92, 0.9, and 0.91, respectively, when no DP is applied. However, at the highest noise level considered (0.9), these metrics drop to 0.85, 0.78, 0.66, and 0.72, respectively. Similar results are seen across the Housing dataset, where predictive performance metrics show a declining trend as the noise level applied increases.

### 5.2. Effectiveness of Differential Privacy in Mitigating MEA

We considering the 3 scenarios of application of DP, namely, DP-Model, DP-Explainer and DP-Model-Explainer, and the baseline No DP scenario. Additionally, when we incorporate DP at explainer, we refer to each case as DP-Explainer-noise level (i.e., DP-Explainer-0.1). This evaluation will allow us to address RQ1 and RQ2. Figures 3 show the *agreement* observed by MEA across the various combinations of applying DP for varying noise levels and number of queries across the *Housing* dataset. We start with No DP (Fig. 3(a)), which allows us to quantify solely the impact of employing different levels of noise at the explainer on the success of the MEA. Results show a general trend where the MEA is more successful as the number of queries used increases across all cases (i.e., independent of the noise level applied). Comparing the *agreement* when employing different noise levels, results show that employing more noise, as expected, provides more defense against MEA. Specifically, with a noise level of 0.9, *agreement* ranges between 50 and 72 when the number of queries increases up to 1000. In contrast, when employing noise levels of 0.5 and 0.1, *agreement* falls within the ranges of 62–75 and 60–78, respectively. In the absence of DP at the explainer, *agreement* starts at 70 with 50 queries and



**Figure 3:** Housing dataset MEA agreement for various DP strategies, noise levels and number queries.

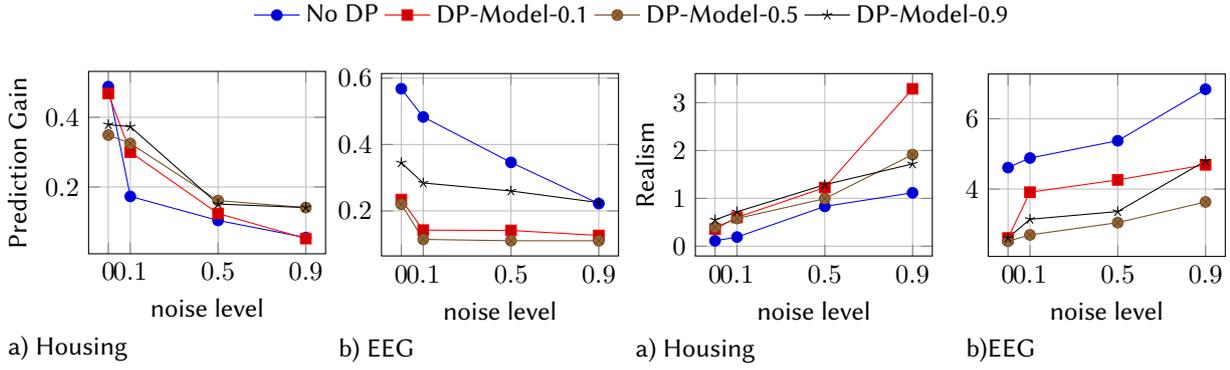


**Figure 4:** EEG dataset MEA agreement for various DP strategies, noise levels and number queries.

reaches 80 when 1000 queries are used. We now focus on the cases where DP is employed at the model level (Fig. 3(b), (c) and (d)). Generally, results show a similar trend across all cases, where *agreement* increases with the number of queries used to perform the MEA. Comparing the *agreement* achieved when employing different noise levels in each case, results show, as expected, that employing higher noise levels at the explainer implies better protection against MEA. For instance, when employing DP-Model with a noise level of 0.1 (Fig. 3(b)), the highest *agreement* observed is 70 when DP-Explainer is employed (which is a DP-Model-Explainer case), compared to 76 without DP-Explainer. Similarly, with a noise level of 0.5 at the model (Fig. 3(c)), the agreement consistently remains lower than in the No DP case, reaching a maximum of 70.63 versus 75 to when DP is only applied at the model. Similar trends were observed to DP-Model-0.9. Figure 4 show the agreement observed by MEA on the EEG dataset across the various cases. The results show similar trends to the one observed with the Housing Dataset. When no DP is applied to the model (Fig. 4(a)), the *agreement* improves with more queries ranging between 68% and 96%, with the highest agreement observed when DP is not applied at all.

### 5.3. Impact of Differential Privacy on Quality of Explanations

Figure 5 shows the prediction gain achieved by explainer across the Housing and EEG datasets for varying noise level. In the Housing dataset, a clear trend emerges as the DP-Explainer noise increases the prediction gain decreases, which means that employing more noise decreases the CF probability toward the desired class. For example, for No DP, the prediction gain starts at 0.488. However, for DP-Explainer with noise level of 0.9 is applied, it drops dramatically to 0.055. This decline is observed consistently across all model noise levels. Moreover, for DP-Model noise levels (0.5 and 0.9) are introduced, the prediction gain prediction observed is less than that of no DP and DP-model 0.1, regardless of the DP-Explainer noise. Similarly, the EEG dataset follows a comparable pattern. In scenarios without DP applied to the model, the prediction gain is lower and ranges from 0.568 to 0.222 as the DP-Explainer



**Figure 5:** Prediction Gain and Realism achieved by explainers in the Housing and EEG datasets under various DP-Explainer noise levels.

noise is higher. When the model is subjected to DP noise at levels of 0.1, 0.5, and 0.9, the prediction gains are consistently lower. We now focus on analyzing the impact of incorporating DP on realism. Across both datasets, increasing the DP-Explainer noise consistently results in higher realism scores, indicating less realistic CFs and degradation in CF quality. In the Housing dataset, even without any DP-model noise, the realism score ranges from 0.113 to 1.116 at a DP-explainer noise of 0.9. This degradation is further amplified when additional DP-Model noise is introduced, e.g. with a DP-Model noise of 0.1, the realism score ranges from 0.356 to 3.289 as DP-explainer noise is higher, and similar patterns are observed for DP-Model-0.5 and 0.9. The EEG dataset exhibits a comparable pattern, although the No DP realism scores are generally higher.

**Discussion on Performance-Privacy-Explanations Interplay:** Results indicate that introducing DP mechanisms affects model performance, although the extent of this impact varies according to the specific use case and dataset. Similarly, the quality of the generated CF explanations is influenced by the privacy parameters applied. Experiments reveal that even slight amounts of noise, whether introduced at DP-Model or within the DP-Explainer, can alter CF quality. In terms of the effectiveness of DP interventions in the context of MEA. Analysis shows that introducing minimal noise at the model level generally offers resistance to MEA. In contrast, higher noise levels provide a more robust defense, albeit at the cost of reduced model performance. When examining the impact of noise on the CFs, we observe that small increments in noise can slightly reduce the success rate of MEA, but further increases yield a more pronounced protective effect. Notably, when both the model and the explainer are simultaneously subjected to DP, a synergistic improvement in resistance to MEA is observed.

## 6. Conclusion

In this work, we investigate the impact of differential privacy (DP) in mitigating model extraction attacks (MEAs) that leverage counterfactual explanations (CFs) within Machine Learning as a Service (MLaaS) environments. We evaluate employing DP implemented at the ML model level via DP-Stochastic Gradient Descent and at the explanation level, and at both simultaneously, to investigate their respective impacts on MEA resilience. Our analysis, conducted across two datasets, demonstrates and quantifies a fundamental trade-off between privacy protection and utility. The introduction of DP noise presents a clear trade-off, as it effectively hinders an adversary’s ability to reconstruct the target model, yet it simultaneously compromises both model performance and the quality of generated CF. Further research will include testing other DP-based methods to generate CFs, MEA methods and more datasets.

## References

- [1] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart, Stealing machine learning models via prediction APIs, in: 25th USENIX Security Symposium (USENIX Security 16), USENIX Association, 2016, pp. 601–618.

- [2] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, 2017, pp. 3–18. URL: <https://ieeexplore.ieee.org/document/7958568>. doi:10.1109/SP.2017.41.
- [3] C. Dwork, Differential privacy, in: International colloquium on automata, languages, and programming, Springer, 2006, pp. 1–12.
- [4] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308–318.
- [5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Computing Surveys (CSUR)* 51 (2018) 1–42.
- [6] F. Ezzeddine, Privacy implications of explainable ai in data-driven systems (2024).
- [7] R. Shokri, M. Strobel, 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.
- [8] C. N. Spertalis, T. Semertzidis, P. Daras, Balancing xai with privacy and security considerations, in: European Symposium on Research in Computer Security, Springer, 2023, pp. 111–124.
- [9] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, *Harv. JL & Tech.* 31 (2017) 841.
- [10] F. Ezzeddine, O. Ayoub, S. Giordano, Knowledge distillation-based model extraction attack using private counterfactual explanations, *arXiv preprint arXiv:2404.03348* (2024).
- [11] U. Aïvodji, A. Bolot, S. Gambs, S. Mehnaz, R. Yvinec, Model extraction from counterfactual explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 99–109.
- [12] W. Abbasi, P. Mori, A. Saracino, Further insights: Balancing privacy, explainability, and utility in machine learning-based tabular data analysis, in: Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–10.
- [13] A. C. Oksuz, A. Halimi, E. Ayday, Autolytus: Exploiting explainable artificial intelligence (xai) for model extraction attacks against interpretable models, *Proceedings on Privacy Enhancing Technologies* (2024).
- [14] A. Yan, R. Hou, X. Liu, H. Yan, T. Huang, X. Wang, Towards explainable model extraction attacks, *International Journal of Intelligent Systems* 37 (2022) 9936–9956.
- [15] Y. Wang, H. Qian, 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.
- [16] P. Dissanayake, S. Dutta, Model reconstruction using counterfactual explanations: A perspective from polytope theory, *Advances in Neural Information Processing Systems (NeurIPS)* (2024).
- [17] S. An, Y. Cao, Counterfactual explanation at will, with zero privacy leakage, *Proceedings of the ACM on Management of Data* 2 (2024) 1–29.
- [18] F. Yang, Q. Feng, K. Zhou, J. Chen, X. Hu, Differentially private counterfactuals via functional mechanism, *arXiv preprint arXiv:2208.02878* (2022).
- [19] S. Pentyala, S. Sharma, S. Kariyappa, F. Lécué, D. Magazzeni, Privacy-preserving algorithmic recourse, *CoRR* (2023).
- [20] D. Nemirovsky, N. Thiebaut, Y. Xu, A. Gupta, CounterGAN: Generating counterfactuals for real-time recourse and interpretability using residual GANs, in: *Uncertainty in Artificial Intelligence*, PMLR, 2022, pp. 1488–1497.
- [21] Scikit-learn Developers, California housing dataset, 2024. URL: [https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\\_california\\_housing.html](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html), accessed: 2024-01-04.
- [22] O. Roesler, Eeg eye state, UCI Machine Learning Repository, 2013. URL: <https://doi.org/10.24432/C57G7J>. doi:10.24432/C57G7J, accessed: 2024-01-04.