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

Fatima Ezzeddine[1], Omran Ayoub[2], Silvia Giordano[2]
[1] Università della Svizzera italiana, Lugano, Switzerland
[2] University of Applied Sciences and Arts of Southern Switzerland

Contribution

In this work, our contribution is:

1. **Novel Model Extraction Attack (MEA):** We propose a novel knowledge distillation (KD)-based MEA that exploit counterfactual explanations (CFs). We simulate an adversarial scenario where an attacker exploit CFs given by a Machine Learning as a Service (MLaaS).
2. **Privacy-Preserving CFs:** We introduce a novel technique to enhance the privacy of CFs generated by GAN-based models. We incorporate differential privacy (DP) into the GAN pipeline. We aim to mitigate the risk of privacy breaches while still providing meaningful explanations.

Specifically, we quantify:
1. The effectiveness of KD-based MEA using agreement metrics.
2. The quality of CFs generated using well-known metrics.

Results demonstrate that:
1. Our proposed KD-based MEA outperforms the baseline.
2. Our proposed private CFs method effectively preserves privacy against MEA while guaranteeing a specific level of privacy and CF quality.

Transparency in MLaaS

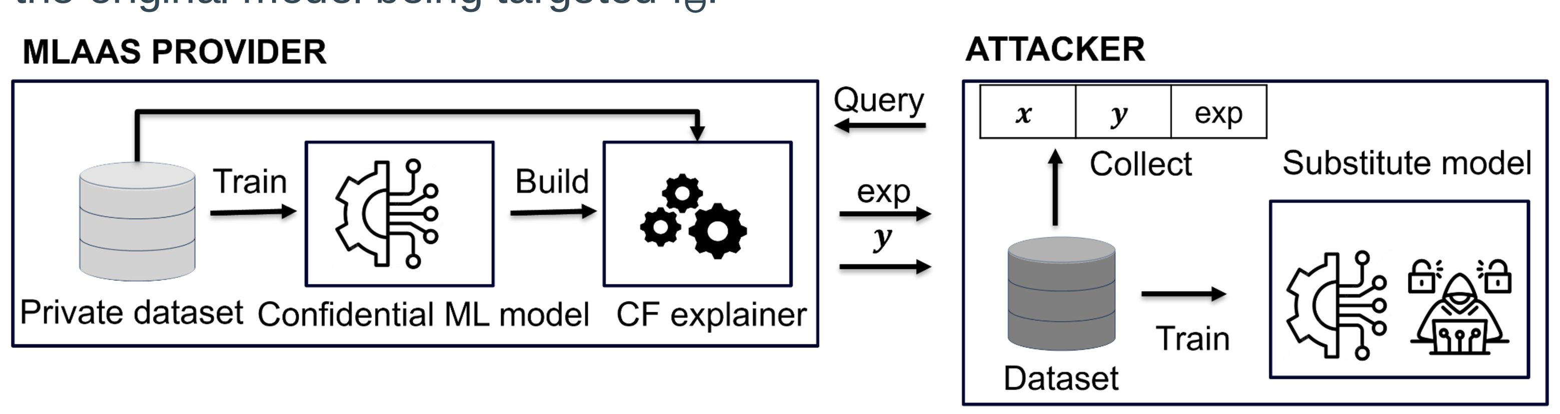
There is a notable increase in the deployment **of MLaaS** across production software applications.

ML models, demonstrably powerful, suffer from lack of **interpretability**. The absence of **transparency**, often referred to as the black box, undermines trust and urges the need for efforts to enhance their explainability.

MLaaS platforms now offer explanations alongside the ML prediction outputs and has elevated concerns regarding privacy, particularly in relation to **privacy leakage attacks** such as model extraction attacks (MEA).

Model Extraction Attack (MEA)

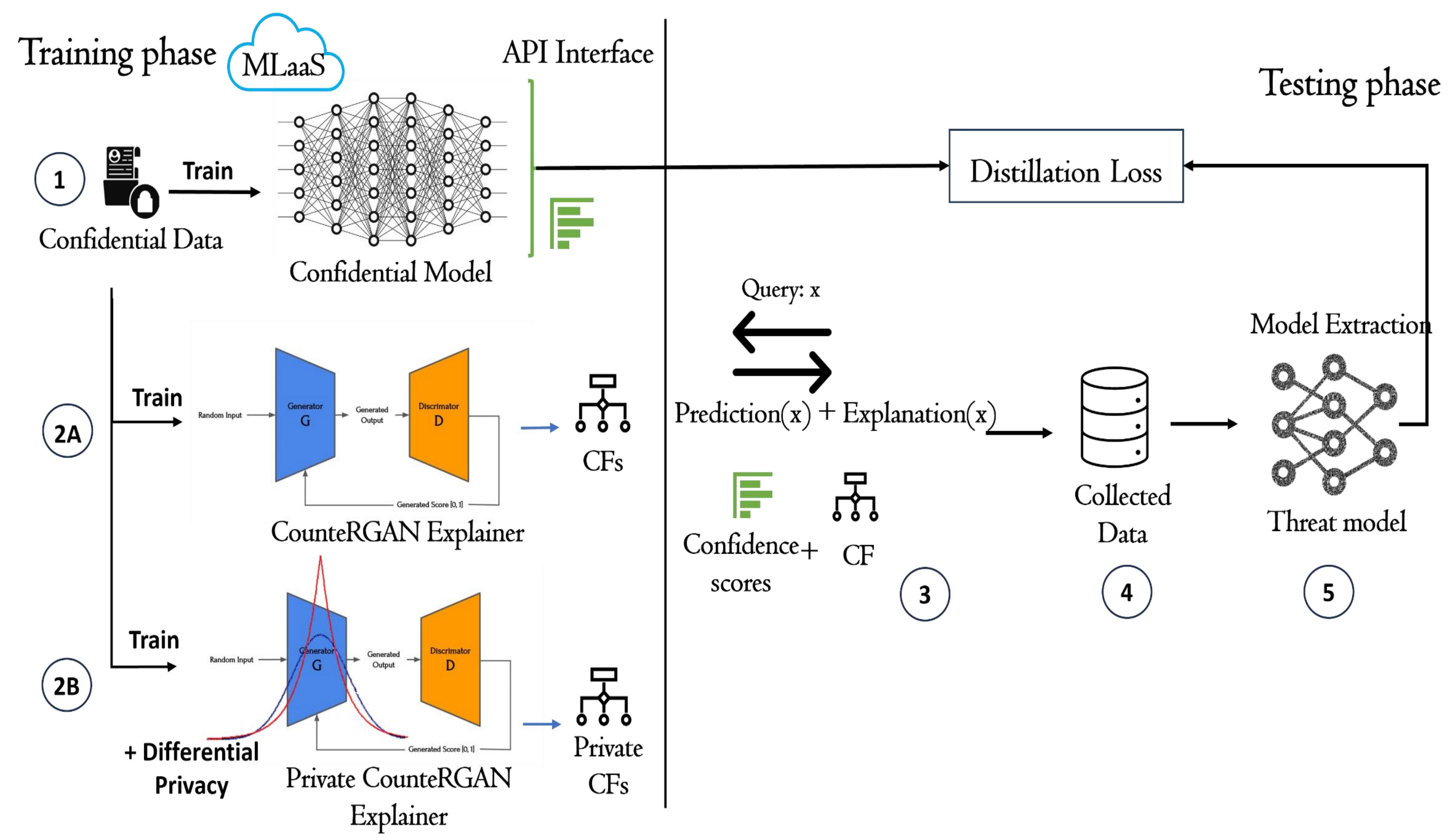
MEA derives a threat model t_{χ} that closely resembles, in terms of functionality, the original model being targeted f_{Θ}.



A user sends a **query**, which includes input data describing a data record x. Once the MLaaS API receives the user query, it performs the prediction $f_{\Theta}(x)$ and generates a CF c=E(x) where $f_{\Theta}(c)$ has a different prediction and returns it along with its corresponding output to the user.

The attacker **extract** (**steal**) f_{Θ} by training a **substitute** (**threat**) model t_{χ}.

Knowledge Distillation-Based MEA



As an owner:
- **Step1:** train a DNN classifier f_{Θ} on a private dataset.
- **Step 2**: train **CounterGAN[1]** CF explainer and deploys as MLaaS.

As an attacker :
- **Step 3**: query the model with **random queries** (with the assumption that the attacker does not have previous knowledge of the training set)
- **Step 4:** collect a dataset to serve as input data to train t_{χ} .
- **Step 5:** apply **KD and** train t_{χ} by minimizing the loss constituted by the threat model **classification loss** in addition to the **distillation loss**. We emphasize the importance of mimicking the output **probabilistic distribution** from the target model to the threat model.

$$loss = \alpha \cdot student\_loss + (1-\alpha) \cdot distillation\_loss$$

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

Counterfactual Explanations generation with Differential Privacy

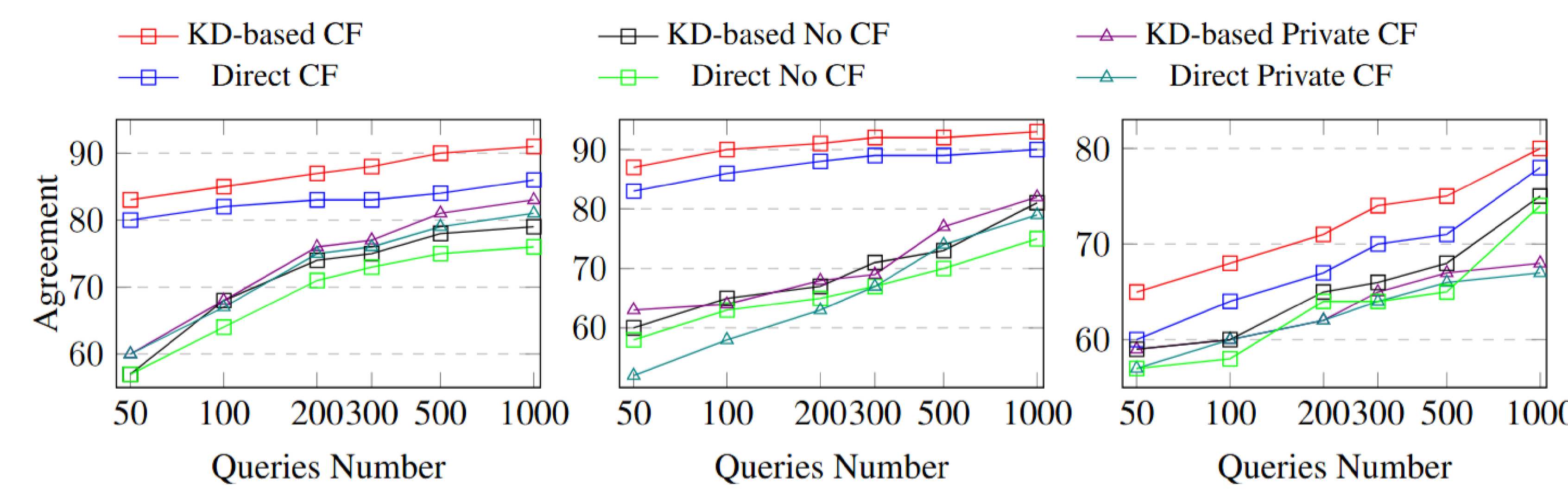
Objective: Prevent the generation of CFs that closely resemble the private training data, reducing the resemblance of CFs with the training data and mitigate the risks of privacy attacks.

Method: Inject DP into the generator during the optimization process. We employ the Adam DP optimizer. The process of DP Adam often involves multiple iterations of **adding noise** repeatedly over several rounds into the **gradients** of the **generator** in addition to a step of gradient **clipping**.

Results and Evaluation

Effectivenes of KD-based MEA

Agreement Metric: Predictions similarity between the ML models t_{χ} and f_{Θ} .



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

Main takeaways:

1. Our proposed KD-based MEA **outperform** the **baseline** by achieving superior agreement levels.
2. When CFs are employed, MEA requires significantly **few instances** to reach a specific high agreement level.
3. The incorporation of **DP** can play a crucial role in **maintaining agreement** levels comparable to scenarios without CFs.

Impact of Incorporating DP in CF generator on quality of explanations

Metrics:
1. **Actionability**: Amount of modification of CFs.
2. **Realism:** If a data instance fits a known data distribution (reconst error of AE).
3. **Prediction Gain:** the probability changes of the CF explanation for a target class

Data	Prediction Gain		Actionability	
	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

Data	Realism		
	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

Main takeways:

1. The initial random query points are inherently unrealistic (do not exhibit high realism).
2. The private CF generation approach ensures this unrealistic nature is preserved when queried with a random data point.
3. The integration of DP has an impact on prediction gain and actionability, constraining the explainer's progress toward the desired class.

References

Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, Abhishek Gupta. Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, PMLR 180:1488-1497, 2022.