Privacy Implications of Explainable AI in Data-Driven Systems

Fatima Ezzeddine[1,2]

[1]Università della Svizzera italiana, Lugano, Switzerland
[2]Scuola universitaria professionale della Svizzera italiana, Lugano, Switzerland

Abstract

Machine learning (ML) models, demonstrably powerful, suffer from a lack of interpretability. The absence of transparency, often referred to as the black box nature of ML models, undermines trust and urges the need for efforts to enhance their explainability. Explainable AI (XAI) techniques address this challenge by providing frameworks and methods to explain the internal decision-making processes of these complex models. Techniques like Counterfactual Explanations (CF) and Feature Importance play a crucial role in achieving this goal. Furthermore, high-quality and diverse data remains the foundational element for robust and trustworthy ML applications. In many applications, the data used to train ML and XAI explainers contain sensitive information. In this context, numerous privacy-preserving techniques can be employed to safeguard sensitive information in the data, such as differential privacy. Subsequently, a conflict between XAI and privacy solutions emerges due to their opposing goals. Since XAI techniques provide reasoning for the model behavior, they reveal information relative to ML models, such as their decision boundaries, the values of features, or the gradients of deep learning models when explanations are exposed to a third entity. Attackers can initiate privacy breaching attacks using these explanations, to perform model extraction, inference, and membership attacks. This dilemma underscores the challenge of finding the right equilibrium between understanding ML decision-making and safeguarding privacy.

Keywords

Explainable Artificial Intelligence, Privacy-Preserving Machine Learning, Privacy Attacks

1. Context and Motivation

In recent years, advancements in Artificial Intelligence (AI) have expanded beyond the primary objective of predictive capabilities. Although accurate predictions are crucial, an equally important goal has emerged: ensuring explainability. Explainability in Machine Learning (ML) models has become a critical objective for making clear and justifiable predictions, especially in high-stakes social decisions. It is essential for these models to offer clear and comprehensible reasons for their predictions and decisions [1]. In this context, Explainable AI (XAI) has emerged as a crucial field of investigation. XAI methodologies are specifically designed to unveil the decision-making processes of complex, opaque models, often referred to as black boxes. With the use of XAI techniques, researchers can gain valuable insights into the reasoning behind model decisions, after they have already been made [2]. XAI techniques employ various methods to interpret the inner workings of complex ML models. These methods generate different types of explanations, e.g., feature importance, counterfactual explanations, etc. To generate tailored explanations, XAI requires a combination of data, interpretable models, and explanatory

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

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

techniques and often incorporates user interaction. Therefore, XAI starts with the foundational element of data, which needs to be diverse and of high quality. This data is not only used to train AI models but also to create explainers. This combination of data, interpretable models, explanatory techniques, and user interaction builds the XAI.

In many applications, the data used to train AI and XAI models contain sensitive information about individuals, such as medical records, or financial transactions, which the GDPR [3] seeks to safeguard. Different approaches are proposed to safeguard sensitive information in data, such as differential privacy (DP) and federated learning (FL). These approaches affect predictive performance to some extent, resulting in a drop in performance, yet they manage to uphold an acceptable level of it. Subsequently, a conflict between ensuring transparency through XAI and ensuring privacy emerges due to their opposing goals. XAI aims to provide insights into model behavior for transparency, while privacy-preserving solutions obscure data to prevent data leakage. Moreover, the output of XAI can unintentionally expose model decision boundaries, leading to potential attacks on privacy [4, 5]. For instance, attackers can exploit XAI explanations such as CFs, which describe the minimal feature value change to alter the model decision and return instances that are close to the decision boundary. FI, which scores the contribution and impact of each feature to the model output exposes information about the gradients in Deep Neural Networks (DNNs) or about the values of the features in ML. In this context, attackers can initiate attacks from these explanations to perform model extraction, inference, and membership attacks [6], especially when the model is shared or deployed publicly on the cloud as ML as a Service (MLaaS). This dilemma underscores the challenge of finding the right equilibrium between explainability and safeguarding private information [4].

2. Background on Explainable Artificial Intelligence

2.1. Motivation and Definition

In order to enhance transparency, XAI techniques provide the necessary tools to open up complex black boxes and shed light on how AI decisions are made [7], promoting fairness, transparency, and accountability within real-world organizations. Moreover, XAI has proven to play a pivotal role in ensuring that AI is trusted and used responsibly. By answering essential "How?" and "Why?" inquiries regarding AI systems, XAI serves as a valuable tool for tackling the increasing ethical and legal issues associated with them. XAI targets diverse entities and includes various stakeholders, such as researchers, model developers like engineers and data scientists, as well as practitioners.

2.2. Post-hoc Explainability

Post-hoc explainability is a technique used to gain insight into the decision-making process of a trained ML model. In this context, post-hoc means that the model's interpretability is addressed after its training, regardless of its complexity or the algorithms used. The approach primarily revolves around the act of querying the model with diverse sets of input data to observe how it reacts to different scenarios. Through these interactions, we can effectively map out the decision boundaries the model uses, shedding light on what factors influence its predictions.

Visualizations and explanations can then be applied to make these insights more accessible and human-friendly, ultimately enabling a better understanding of the model's predictions. These visual aids are essential in making the insights gained more accessible to data scientists, end users, and domain experts who are willing to understand why the model is making specific predictions. By going through this process, post-hoc explainability serves a vital role in improving model transparency and building trust in its performance.

Understanding an AI system with XAI relies on its training data, process, and model. Therefore, XAI can be applied throughout the entire AI development pipeline. Specifically, it can be applied in different stages of modeling, such as before, during, and after (post-modeling explainability). In this work, the primary emphasis will be on post-modeling XAI (Post-hoc), since ML models are often developed with only predictive performance in mind.

Feature Importance Feature Importance (FI) explanations involve assigning a quantitative measure in the form of a numerical score to each input feature within a given model. The primary goal of calculating FI is to discern which features have influential effects on the model's predictions and which ones have a relatively lesser impact. These importance scores help practitioners and data scientists gain insights into which factors are most critical in influencing those decisions. Features that, when modified, cause more substantial shifts in the model's output are considered more important because they have a greater influence on the final prediction. For deep learning models, many feature-based explanation functions are gradient-based techniques that analyze the gradient flow through a model. Approaches such as Layer-wise Relevance Propagation (LRP) [8], and Deep Learning Important FeaTures (DeepLIFT) [9] exist.

Counterfactual Explanations CFs leverage the concept of potential outcomes to assess causal relationships within a data-model framework. CFs empower informed decision-making and the implementation of explainable, accountable, and ultimately more ethically responsible AI [10]. It achieves this by constructing a hypothetical scenario, distinct from the observed data, and evaluating the corresponding model output under this scenario. The generation of informative and interpretable CFs necessitates the optimization of well-defined metrics [11] such as diversity, validity, proximity, and user constraints. Conversely, model-specific methods tailor the cost function optimization process to leverage the inherent characteristics of the employed model. For instance, in the case of differentiable models, gradients play a critical role in guiding the optimization towards finding CFs [12]. Conversely, model-agnostic methods achieve generalization across diverse model architectures [13, 14].

3. Related Work: Interplay between XAI and Privacy

3.1. Context and Problem Formulation

Data protection and *privacy* is one of the primary dimensions in ML and AI. It involves ensuring that the data used to train and test ML models does not expose sensitive information about individuals or entities. This is particularly critical when dealing with datasets that contain personally identifiable information or confidential details. Techniques like anonymization and DP have emerged as valuable tools in the data privacy field. They allow us to protect the privacy of individuals represented in the data, even as we leverage it to train models. Beyond data privacy, *model privacy* is also a pressing concern. The architecture of ML models can be

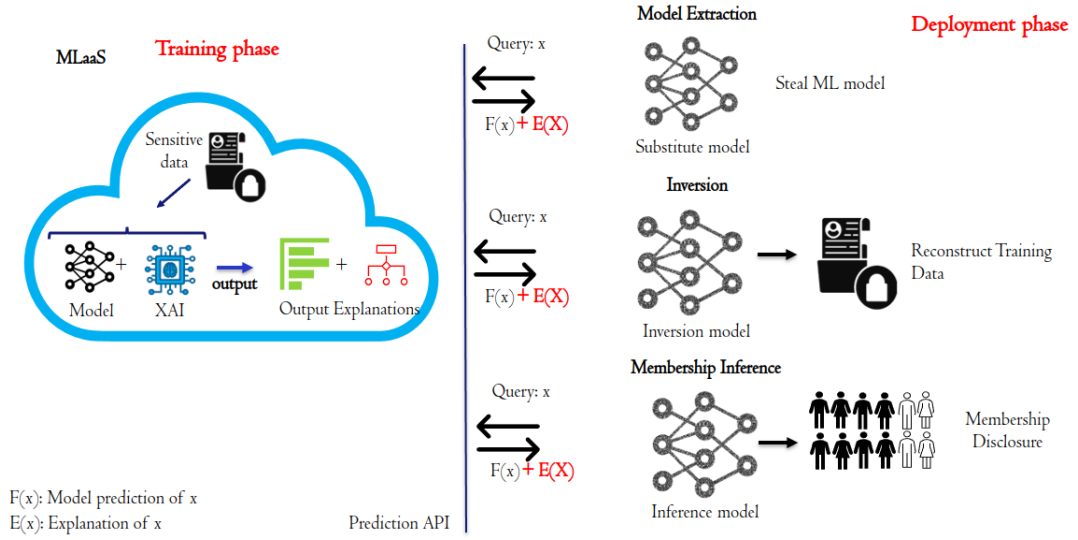

Figure 1: Scenario of privacy attacks where MLaaS provides explanations alongside the prediction

susceptible to privacy breaches. Models may unintentionally encode information about the training data they were exposed to, and this could pose risks when shared or deployed publicly on the cloud as MLaaS. Attacks such as model extraction, inversion, or membership inference can exploit these vulnerabilities (Details in the following sections and Fig. 1). However, privacy is not included in the default behavior of most ML algorithms. They tend to learn not just the general trends but also the specifics of the data, potentially revealing sensitive information when the model is made public. In an ideal scenario, we want these algorithms to focus on extracting general trends and patterns from the data while deliberately avoiding the inclusion of specific details about the data. This emphasis on distilling general patterns means that the algorithms should primarily capture the fundamental, common insights that are valuable for decision-making, aligning with privacy concerns, as they identify important details without risking individual privacy. XAI can inadvertently compromise privacy by revealing sensitive information about the model's decision boundaries. Moreover, the process of returning real data points with CFs can inadvertently expose specific instances from the training set or behaviors. Also, the process of assigning FI scores exposes the values of gradients and the feature values themselves. This conflict makes striking the right balance between model explainability and data privacy crucial to ensuring that XAI enhances our understanding of AI systems without leaking individual privacy.

3.2. Attacks on Machine Learning Models

3.2.1. Membership inference Attacks

A membership inference attack (MIA) is a privacy-related threat in ML where an adversary attempts to determine whether a specific data point was part of the training dataset of a deployed

model [15, 16]. MIA are particularly concerning because they can compromise the privacy of individuals whose data was part of the training dataset. If an attacker can determine that a specific data point was included in the training data, it may reveal sensitive information about that individual, even if the model's output does not directly disclose such information. To perform membership attacks, [15] proposes a shadow training process that mimics the target model with shadow models, and trains the attack model using data that is extracted using data synthesis. Also, [17] discusses and proves that points with a very high loss tend to be far from the decision boundary and are more likely to be non-members. Regarding how explanation can facilitate performing MIA, [4] quantifies information leakage in model predictions when explanations are provided. The authors evaluate feature-based explanations, highlighting how back-propagation-based explanations reveal decision boundaries.

3.2.2. Model Extraction Attack

Model extraction (MEA) is a class of attacks where an adversary tries to reverse-engineer a target model by observing its behavior and querying it. MEA can potentially lead to the theft of intellectual property compromising proprietary models [18, 19]. Authors in [19] discuss the weakness in ML services that take incomplete data with confidence levels and show successful attacks on different ML models like decision trees, SVMs, and DNNs by using equation-solving, path-finding algorithms. Regarding how explanations can facilitate MEA, FIs, and CFs have proved their ability to reveal the decision boundary of a target model [20].

[21] perform the attack by minimizing task-classification loss and task-explanation loss. Authors in [22] show how gradient-based explanations quickly reveal the model itself and highlight the power of gradients. Regarding CFs, [23] proposes a strategy to target the decision boundary shift by taking not only the CF but also the CF of the CF as pairs of training samples.

3.2.3. Model Inversion Attack

A model inversion attack (MINA) is a privacy-related threat in ML where an adversary attempts to reconstruct sensitive or private information about individual data points from trained model predictions. In other words, the MINA task is to predict the input data, that is, the original dataset for the target model. In [24] discusses how providing explanations harms privacy and studies this risk for image-based MINA on private image data from model explanations. The authors developed several CNN architectures that achieve significantly higher inversion performance than using only the target model prediction. To minimize the risk of MINA, [25] presents a generative noise injector for model FI explanations by perturbing model explanations.

4. Research Questions and Objectives

We pose the following research questions (RQs):

1. To what extent does the utilization of known privacy-preserving techniques, such as DP, effectively safeguard privacy and prevent information leakage when combined with explanations provided by XAI?

2. Can we produce high-quality XAI explanations while safeguarding privacy to mitigate potential vulnerabilities to attacks?
3. Which approach, privacy-preserving XAI or privacy-preserving ML, offers a more effective solution for safeguarding sensitive information in XAI systems?

To address RQ1, we aim to evaluate the trade-off and assess the effectiveness of existing privacy-preserving techniques (e.g., DP) in mitigating information leakage when combined with XAI explanations for CFs and FI. This will involve investigating the extent to which explanations can be exploited for privacy attacks like MIA, MEA, or MENA.

To address RQ2, we aim to explore the possibility of generating high-fidelity XAI explanations while simultaneously safeguarding privacy.

Such approaches aim to develop an XAI framework that concurrently optimizes two objectives: *i) generating high-quality CFS*, and *ii) adhering to pre-defined privacy constraints*. Furthermore, the integration of DP during the backpropagation of gradients for FI computation is another promising avenue for investigation.

To address RQ3, we will conduct a comparative analysis of privacy-preserving XAI and privacy-preserving ML techniques. This analysis will evaluate their strengths and weaknesses in safeguarding sensitive information within XAI systems. By comprehensively assessing these aspects, we aim to identify the approach that offers a more robust and enduring mechanism for privacy protection within XAI applications, covering different types of data.

5. Results and contributions to date

In the initial research, I explored CF generation through RL, with the specific goal of constructing an explainer that operates independently of input data. The investigation then progressed to a more in-depth examination of CFs, focusing on their potential for information leakage and their ability to reveal the decision boundaries of ML models. To reach this aim, a new methodology is proposed to carry out MEA through a concept known as knowledge distillation (KD). I also delved into the domain of explainable deep learning methods within distributed systems, such as Vertical Split Learning (VSL), aiming to evaluate the potential information disclosure resulting from FI across various entities. In addition, I analyzed the impact of DP on the explainability of anomaly detection (AD) models. More specifically:

1. **Explored how RL can be leveraged to generate CF explanations** without relying on the dataset as input to the explainer. The main aim is to let the CF generator learn generalizable patterns from the training data without exposing it. The explainer determines which features to modify and by how much, by maximizing a custom reward function designed to jointly optimize various metrics.
2. **Designed a new attack approach to evaluate the use of KD for an MEA** in scenarios where CFs are given to an attacker. I benefit from the property of KD and the process of transferring knowledge from a large model to a smaller one. The findings reveal that employing KD with the presence of CFs can indeed yield successful MEA.
3. **Proposed an approach to generate private CFs** I introduce the concept of DP within the GANs CF generation pipeline to generate CFs that deviate from the statistical properties of the confidential dataset, offering a layer of protection against potential privacy breaches.

4. **Explored VSL** strategies and performed experiments to explore the risk of information leakage regarding the original features using gradient-based explanations (IG and DeepLIFT). My application of VSL focused on a use case related to Network Function Virtualization. My findings highlight how an attacker on the server side can exploit XAI techniques to achieve additional tasks, without access to the original features.
5. **Explored DP with AD** Analyzed the trade-off between privacy achieved by DP and explainability achieved using SHAP.

6. Expected next steps and final contribution to knowledge

This PhD research aims to achieve significant advancements in bridging the critical gap between XAI and data privacy. We will address the inherent conflict between providing users with clear explanations of AI models and protecting their sensitive data (privacy). We aim to develop a defense mechanism in the form of high-quality explanations while simultaneously ensuring privacy.

References

[1] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57.

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

[3] P. Regulation, General data protection regulation, Intouch 25 (2018) 1–5.

[4] 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.

[5] S. Goethals, K. Sörensen, D. Martens, The privacy issue of counterfactual explanations: explanation linkage attacks, ACM Transactions on Intelligent Systems and Technology 14 (2023) 1–24.

[6] M. Rigaki, S. Garcia, A survey of privacy attacks in machine learning, ACM Computing Surveys (2020).

[7] T. Speith, A review of taxonomies of explainable artificial intelligence (xai) methods, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2239–2250.

[8] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

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

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

[11] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: A review, arXiv preprint arXiv:2010.10596 (2020).

[12] P. Wang, N. Vasconcelos, Scout: Self-aware discriminant counterfactual explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8981–8990.

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

[14] T. M. Nguyen, T. P. Quinn, T. Nguyen, T. Tran, Counterfactual explanation with multi-agent reinforcement learning for drug target prediction, arXiv preprint arXiv:2103.12983 (2021).
[15] 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.
[16] N. Patel, R. Shokri, Y. Zick, Model explanations with differential privacy, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1895–1904.
[17] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, H. Jégou, White-box vs black-box: Bayes optimal strategies for membership inference, in: International Conference on Machine Learning, PMLR, 2019, pp. 5558–5567.
[18] S. Pal, Y. Gupta, A. Shukla, A. Kanade, S. Shevade, V. Ganapathy, Activethief: Model extraction using active learning and unannotated public data, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 865–872.
[19] 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), 2016, pp. 601–618.
[20] T. Miura, S. Hasegawa, T. Shibahara, Megex: Data-free model extraction attack against gradient-based explainable ai, arXiv preprint arXiv:2107.08909 (2021).
[21] 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.
[22] S. Milli, L. Schmidt, A. D. Dragan, M. Hardt, Model reconstruction from model explanations, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 1–9.
[23] 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.
[24] X. Zhao, W. Zhang, X. Xiao, B. Lim, Exploiting explanations for model inversion attacks, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 682–692.
[25] H. Jeong, S. Lee, S. J. Hwang, S. Son, Learning to generate inversion-resistant model explanations, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 17717–17729. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/70d638f3177d2f0bbdd9f400b43f0683-Paper-Conference.pdf.