Evaluating the Privacy Exposure of Interpretable Global and Local Explainers

Francesca Naretto[*], **Anna Monreale**[*], **Fosca Giannotti**[**]

[*]University of Pisa, Italy

[**]Scuola Normale Superiore, Italy

E-mail: francesca.naretto@unipi.it,
anna.monreale@unipi.it, fosca.giannotti@sns.it

Received 6 March 2024; received in revised form 15 October 2024; accepted 17 November 2024

Abstract. During the last few years, the abundance of data has significantly boosted the performance of Machine Learning models, integrating them into several aspects of daily life. However, the rise of powerful Artificial Intelligence tools has introduced ethical and legal complexities. This paper proposes a computational framework to analyze the ethical and legal dimensions of Machine Learning models, focusing specifically on *privacy* concerns and *interpretability*. In fact, recently, the research community proposed privacy attacks able to reveal whether a record was part of the black-box training set or inferring variable values by accessing and querying a Machine Learning model. These attacks highlight privacy vulnerabilities and prove that GDPR regulation might be violated by making data or Machine Learning models accessible. At the same time, the complexity of these models, often labelled as "black-boxes", has made the development of explanation methods indispensable to enhance trust and facilitate their acceptance and adoption in high-stake scenarios.

Our study highlights the trade-off between interpretability and privacy protection. By introducing REVEAL, this paper proposes a framework to evaluate the privacy exposure of black-box models and their surrogate-based explainers, whether local or global. Our methodology is adaptable and applicable across diverse black-box models and various privacy attack scenarios. Through an in-depth analysis, we show that the interpretability layer introduced by explanation models might jeopardize the privacy of individuals in the training data of the black-box, particularly with powerful privacy attacks requiring minimal knowledge but causing significant privacy breaches.

1 Introduction

Recent developments have led to the widespread integration of Artificial Intelligence (AI) systems into diverse aspects of our everyday routines. While this phenomenon may appear mostly positive, it also calls for ethical discussions and regulatory measures to ensure a responsible usage of these systems. In fact, the availability of Big Data is bringing us into a new era in which decisions are being made based on the knowledge distilled from digital traces generated by the use of digital tools that are now present in everyday life. These traces are being collected and analyzed at individual, group, and societal levels, allowing for the development of powerful AI systems that can be used in critical domains such as medicine, finance or autonomous vehicles. This setting poses several challenges to the respect of ethical values such as interpretability, privacy and accountability. In fact, AI systems exploit sensitive user data for their training. This requires collecting, storing and exchanging data: during all these passages, the protection of personal data must be taken into account. In this context, we can find some well-known accidents, such as the Cambridge Analytica scandal and the Uber data breach, where unintended disclosure of personal data has resulted in harm to individuals[1]. In addition, the employment of sensitive data to train AI systems poses privacy concerns despite the fact that the data is kept private. In fact, the AI models deployed learned crucial patterns and information from the private data. Because of this vulnerability, AI systems based on Machine Learning models (ML) are vulnerable to various privacy attacks, such as the Model Inversion Attack and the Membership Inference Attack, which can infer the data used to train the model simply by querying the model itself. In recent times, the number of privacy attacks of this type has increased considerably, with different variants of these attacks having different underlying assumptions[8, 40, 28]. For these reasons, nowadays it is of crucial importance to address the data privacy aspect also when only the AI system is published.

The importance of data privacy in AI applications is further highlighted by the European Union's introduction of the General Data Protection Regulation (GDPR) in 2018, which establishes rules for companies' use and management of personal data. The GDPR and similar legal frameworks aim to regulate data breaches and minimize the harm caused to individuals and organizations. Ensuring users' privacy in the training set is just one concern posed by AI systems. In fact, we need to consider that these systems are often based on complex ensemble models and Neural Networks (NN) that are referred to as "black-boxes" due to their opaque internal structure and decision-making process. This lack of transparency and interpretability can limit the trust in these systems, especially in high-stakes decision-making. The need for an explanation is also referred to in the GDPR, as well as being listed as a crucial requirement for having a trustworthy AI system in the Assessment List for Trustworthy Artificial Intelligence (ALTAI) document and in the novel proposal of the Artificial Intelligence Act (AIA)[2]. To address the need for an explanation, the eXplainable Artificial Intelligence (XAI) literature has developed two families of explainers: local explainers, which explain the reason for a specific instance classification, and global explainers, which explain the logic of the ML model as a whole. In [24], it has been shown that the layer of interpretability added by an interpretable model may jeopardize the privacy protection of individuals represented in the data used for training a black-box classifier. In particular, in that setting the authors considered global explainers as learned functions derived by exploiting the predictive knowledge of a black-box model learned on a private dataset. In that setting, they proved that attacking the privacy of the explainers leaks more information with respect to their black-boxes, leading to a higher privacy exposure.

Recent research has explored the privacy concerns associated with explainers. Notably, in [37], the authors examined the privacy implications of explanations based on backpropagation, leveraging gradients, such as GradCam [36], as well as perturbation-based methods, like SmoothGrad [39] and LIME [33]. Due to the structure of the explanations, these methods focus on NN. Their analysis discovered that backpropagation-based explanations pose privacy risks, particularly for minority groups, while LIME and SmoothGrad do not. Furthermore, [32] explored this issue in the context of images. Their study assessed the effectiveness of Membership Inference Attacks and Evasion Attacks against various explainers, including the ones already mentioned.

In contrast to the existing literature, our work introduces a novel framework named RE-

[1] Cambridge Analytica Scandal, Uber data breach

[2] ALTAI, AIA

VEAL (pRivacy risk EValuation of Exposing surrogAte expLainers). This framework systematically evaluates the privacy risk associated with black-box models and their explainers, whether they operate *locally* or *globally*, using surrogate models. Notably, REVEAL is agnostic to the black-box's structure and is versatile in accommodating different privacy attack strategies and surrogate-based explainers. The primary objective of REVEAL is to identify any alterations in privacy exposure that may arise when disseminating the black-box model and/or its explainers. Differently from other works from the literature, we not only explore the effect of the most popular privacy attack against black-box models, namely the Membership Inference Attack, but we also analyze other privacy attacks, which require fewer assumptions and background information with respect to the original privacy attack and hence are more powerful and pose a greater risk to users' privacy.

The remaining of the paper is organized as follows: in Section 2, we present the literature related to XAI and to the Privacy, focusing on privacy breaches in ML models. Then, Section 3 introduces the main problems we tackle in this paper, as well as the basic notions useful for understanding our proposal. Section 4 describes the main steps of REVEAL, our assessment methodology for global and local explainers, while Section 5 discusses the experimental results obtained by applying our framework, exploring the privacy exposure obtained in various settings. Lastly, Section 6 concludes this paper by highlighting the main findings of the work and possible future directions.

2 Related Work

This paper presents a methodology for assessing the privacy risk of black-box models and their explainers, based on surrogate models, being them global or local. Accordingly, in this section we first review the literature related to privacy, and in particular we will focus on the task of assessing the privacy risk, then to XAI.

2.1 Privacy

Privacy has been a subject of concern in various fields, with two main objectives. First, the primary goal is to evaluate the privacy risks of people involved in information systems of various kinds, including also ML tasks. Once the privacy risk has been assessed, there is the need to shelter information systems against harmful disclosures of sensitive information. In the following, we describe in detail the first procedure, e.g. the *privacy risk assessment*, given the topic of this work. The first objective of data privacy is to evaluate the privacy risks of users represented in a dataset using a privacy risk assessment methodology. Depending on the results of this assessment, a privacy protection technique can be applied to data or ML models to safeguard users from malicious adversaries. These protection techniques are based on established privacy models such as randomization, differential privacy, and k-anonymity [44, 35, 11, 9]. They apply some transformation on the data or the ML models in a way that guarantees specific thresholds on the risk of privacy leaks. However, the main objective of our work is to assess the privacy risk. For this reason, in the following we present the literature related to the field of *privacy risk assessment*.

Assessing the privacy risk requires quantifying the release of sensitive information, which can occur by accessing data directly [44] or by accessing ML models [1, 38, 12]. The problem of privacy disclosure through ML models is a recent breakthrough. In fact, ML models learn from data, and even if the data is not exposed, querying the model may still lead to privacy leaks about the individuals in the training dataset. In the following, we first describe the main procedure to assess the privacy risk analyzing the data and then we focus on the privacy risk assessment of ML models.

In the context of privacy risk assessment on the data, Pratesi et al. proposed PRUDEnce[30]: a framework enabling a systematic assessment of empirical privacy risk concerning specific privacy attacks on data. Technically, the framework simulates the presence of an adversary that tries to re-identify the people in the dataset under analysis. To this end, PRUDEnce generates all the possible background knowledge about the users of the dataset that the adversary may know, and assesses the risk with respect to the worst case scenario. This methodology is general and allows for a privacy risk assessment of different kinds, depending on the kind of data and privacy attacks considered. In recent years, the evaluation of privacy risks in datasets has gained significant attention, with research exploring various directions. A key focus has been on sequence data, such as trajectories or features extracted from them, which has been extensively analyzed. For instance, in [26, 27, 25], privacy risk evaluation adopts a user-centric perspective, aiming to help users understand why a particular privacy risk has been identified and how it can be mitigated. Similarly, Gomes [16] recently conducted an in-depth analysis of privacy risk assessment specifically for trajectory data. Beyond trajectories, sequential data such as text has also been investigated. In particular, given also the complexity of the data under analysis, there are approaches that focus on analyzing the psychometric profiles of users [22].

In recent years, similar approaches have been proposed to evaluate the privacy exposure of ML models. The primary objective in this context is to determine whether the ML model unintentionally reveals sensitive information. One of the most popular attacks of this kind is the Membership Inference Attack (MIA), proposed by Shokri et al. [38]. In this case, the aim is to infer the membership of a given record to the training set of a classification model. Following this work, Choquette-Choo et al. [8] proposed a variant of the original MIA, called LABELONLY attack, in which some of the assumptions of the Shokri's attack are relaxed. In particular, MIA needs the probability vector for inferring the membership of a record while LABELONLY exploits only the hard labels. Recently, Rizzo et al. proposed ALOA ([23]), a variant of the LABELONLY, which assumes an adversary with weaker prior knowledge with respect to the LABELONLY, i.e. no statistical information are known by the attacker, showing a higher privacy risk exposure. In addition to MIA and its variants, Fredrikson et al.[13, 12] designed the so-called reconstruction attacks, where the attacker's objective is to reconstruct one or more training samples and their respective training labels. Another type of attack is the property inference attack, introduced by Ganju et al. [15], which aims to extract unintentionally learned information that is not explicitly encoded as features in the ML model. For instance, property inference attacks can uncover information such as the gender ratio in the training data set. Such attacks can be used in tandem with MIA or reconstruction attacks to enhance the adversary's knowledge.

2.2 Explainable Artificial Intelligence

Today, interpretability is a crucial area of research, with the objective of explaining the internal reasoning of ML models, commonly known as black-box models, due to their complex internal reasoning that is often difficult to comprehend. A comprehensive analysis of the current state-of-the-art in this field is available in [2]. For this work we focus on *post-hoc* explainers, in which, given a black-box b, the main objective is to explain it from the outside, without modifying the original ML models. In this context, there are two main types of explainers: global explainers, which describe the overall behaviour of the ML model, and local explainers, which aim to explain the internal reasoning of the ML model when classifying a single record. Among the most popular explainers for obtaining local explanations include LIME [33] and SHAP [21], which provide feature importance explanations, e.g. they assign an importance value for each variable in the input record, and LORE [17], which outputs rules and counterfactual rules (i.e. rules to follow to obtain the opposite prediction). For the global case, most of the research in this area focuses on tabular data, where the surrogate models produced may be Decision Tree (DT) [10, 3], rule-based classifiers [34], or prototypes [20]. In this paper, we focus on surrogate-based explainers, and we exploit TREPAN [10] and LORE [17], which outputs tree-based structures and rules. TREPAN is one of the first global explainers proposed in the literature, which aims to explain the overall behavior of a black-box by using an enrichment of the input training data to define a DT global surrogate model. A detailed explanation of TREPAN is provided in Section 3. LORE, instead, is a local explainer which outputs rules and counterfactual rules by exploiting a genetic algorithm to generate synthetic neighbourhoods around the input record.

3 Preliminaries

Before describing the details of our framework, we first present the legal and ethical reasons behind the necessity of creating REVEAL (Section 3.1). After having presented the motivation behind the development of this framework, we then introduce some basic notions that are fundamental for understanding the details of our approach. In particular, we first define what a black-box model is and introduce the nomenclature used throughout this paper in Section 3.2. Following, we describe the surrogate-based explainers employed in our work, beginning with the global explainers in Section 3.3, and then the local explainer in Section 3.4. Lastly, we provide the details of the three privacy attacks against the black-box models we exploit to validate REVEAL: firstly, in Section 3.5, we describe the first version of the Membership Inference Attack (MIA). Following, we introduce two variants of MIA: in Section 3.6 we present LABELONLY, a fast variant of MIA in which the attacker only needs the hard labels out of the black-box, while in Section 3.7 we report ALOA, an agnostic version of MIA, in which the majority of the assumptions of MIA are relaxed.

3.1 Legal framework for AI

In recent years, the increasing adoption of AI systems has prompted the introduction of numerous regulations and laws to govern these systems while promoting their ethical and trustworthy deployment. The primary objective of these diverse laws can be summarized as promoting the advancement in AI technology while mitigating potential risks. Given the multitude of regulations worldwide, various ethical considerations come into play. These include the necessity for data governance and record-keeping to ensure accountability, reliability, accuracy, and resilience of AI systems. In addition, cybersecurity is a key concern cited by the majority of the legislation ([47, 7, 29, 45, 19]). Some laws emphasize the need to foster innovation ([7, 29]), while others concentrate on upholding human rights and democratic values specific to their respective nations ([45, 47]) while promoting innovation and with a clear aim in becoming a leader in the sector.

Despite the disparities among global legislation, they all concur on one crucial aspect: there is the need to achieve better trust in the AI systems and to reach this objective comprehending the rationale behind AI outputs is a mandatory task. This challenge originates from the opacity of ML models employed in AI systems. These models consist of intricate mathematical functions that are often challenging or even impossible to decipher, making them the so-called *black-box* or *opaque* models. The difficulty of understanding the internal workings of ML models leads to several disadvantages. Most notably, it hampers the ability to debug model behaviour and identify errors. These errors may encompass technical misclassifications, but they can also involve biases and discrimination against minority groups. As an example, consider the infamous Compas recidivism case [31], where minorities were wrongly categorized as high-risk individuals, resulting in their unjust incarceration due to historical data that associated them with high-crime neighbourhoods. Given this context, understanding the inner workings of AI systems is crucial for addressing these challenges. While developing appropriate explanations—such as those that are comprehensible, reliable, and faithful to the model—and methods to validate them remains complex, terms like *explanations*, *interpretation*, and *transparency* are key to modern AI regulations, including the AIA and those from China, the UK, the United States, and Japan ([7, 19, 45, 29]).

The United States, for instance, proposed the *Algorithmic Accountability Act* [42] and the more recent *Blueprint for a Bill of AI Rights* [43], in which it is highlighted the paramount importance of transparency and comprehending the decisions made by automated decision-making systems. These documents explicitly mention the need to use Explainable AI techniques and indicate the government's commitment to researching and establishing best practices in this domain. When considering Europe, there is the AIA. It uses a risk-based approach, similar to the *General Data Protection Regulation (GDPR)*: depending on the risk level, there are different requirements. The categories listed in the AIA are *unacceptable* risk, *high* risk, *limited* risk and *minimal* risk, depending on the context of the use of the AI system, the potential impact on the health and safety of people as well as the respect of the fundamental rights of persons. The *unacceptable* risk category is the one with the highest risk: the AI systems that fall into this category are considered a danger to the people who use them, a threat to the safety, livelihoods, and rights of people. Hence, the AIA imposes very strict limitations on this category of systems. Examples of AI systems in this category are social scoring by governments or toys with voice assistance that may encourage dangerous behavior. The rest of the categories have lower risk levels, in decreasing order starting from *high*, *limited*, and *minimal* risk. For *high* risk AI systems, the AIA requires clear and understandable information about their abilities and limitations, as well as transparent decision-making processes. In this category we can find AI systems involved in critical infrastructures, such as transports, in which the life and health of people may be put at risk, or AI models part of the law enforcement or migration.

In the context of high-risk systems, the AIA mentions that the data subjects should have the right to obtain explanations for AI-driven decisions. In particular, the primary concern of the AIA is on transparency and human oversight, including aspects of documentation and risk management, addressing the opacity of AI systems in a holistic manner, as a result of an interrelated set of attributes of the AI system. Even if not directly mentioned, the *high* risk systems align effectively with Explainable AI techniques, which explain the internal reasoning of complex AI models. Providing documentation and human oversight is part of the things to do to achieve transparency, but it may not be sufficient.

Considering the *limited* risk systems, they have less strict requirements, but they still must be transparent, informing users about their capabilities, reasoning, and limitations. In addition to this setting, it is important to mention that also in the General Data Protection Regulation [48], into force since 2018, the topic of interpretability was mentioned, stating that users have the right to an explanation.

When considering the United Kingdom, in its AI *Regulation Policy Paper* [45], it prioritizes Explainable AI as a mandatory technique. In particular, the UK recognizes the importance of ensuring that AI systems are transparent and accountable to the public, while acknowledging the ongoing technical challenges associated with providing comprehensive and stable explanations. Also China, a key player in AI development, addressed the transparency concerns. The *Internet Information Service Algorithmic Recommendation Management Provisions* [6], the nation's first algorithm regulation, emphasizes the significance of Explainable AI in building trust and improving AI model development, as well as the *New Generation Artificial Intelligence Development Plan* [7].

From this first description of the AI regulations around the world, it is evident that there are similarities but also differences among them. In [18] the authors first analyze the similarities and differences between the AIA and US Algorithmic Accountability Act. In particular, the authors highlight how the former is focused on a long-term plan, in which the interest is to recreate a Bruxells effect, as in the case of GDPR. In the case of the US, on the other hand, regulation is sound and consistent, but focused more on the present than the long term. In particular, the authors also address the issue of a possible Washington effect.

On a global scale, UNESCO's *Ethics of Artificial Intelligence* publication [46] emphasizes several key principles, with transparency as a central focus. This international organization highlights transparency's crucial role in fostering the responsible and ethical use of AI technologies all over the world. These legislative efforts reflect a shared consensus that transparency and a deep understanding of AI systems are essential for achieving trustworthy AI systems, despite the technical challenges of providing comprehensive explanations.

Even if there are several advantages in using XAI, it can also pose privacy issues: the explainability techniques often need more detailed data to provide reliable explanations, which can be a concern, especially when dealing with sensitive information. When dealing with this kind of problem, we refer to the GDPR, a European regulation since 2018, which requires assessing the privacy risks for individuals involved and, based on the results of the assessment task, it requires protecting the privacy of the users under analysis. Given the interplay among the AIA, GDPR, privacy, and XAI, we saw the need for the definition of REVEAL: a framework able to assess the privacy exposure of the AI model and its explainer.

3.2 Black-box models

A classifier, is a function $b : \mathcal{X}^{(m)} \rightarrow \mathcal{Y}$ which maps data instances (tuples) x from a feature space $\mathcal{X}^{(m)}$ with m input features to a decision y in a target space $\mathcal{Y}$ of size $L = |\mathcal{Y}|$, i.e., y can assume one of the L different labels ($L = 2$ is binary classification, $L > 2$ is multi-class classification). We use $b(x) = y$ to denote the decision y taken by b, and $b(X) = Y$ as a shorthand for $\{b(x) \mid x \in X\} = Y$. Instead, we denote by $\overline{y}_b$ the probability vector of size L in which the sum of all the values is one. An instance x consists of a set of m attribute-value pairs (a_i, v_i), where a_i is a feature (or attribute) and v_i is a value from the domain of a_i. The domain of a feature can be continuous or categorical. We assume that a classifier is available as a function that can be queried at will. In case b is a complex classifier, whose internals are either unknown or known but uninterpretable by humans, it is called *black-box* classifier. Examples of black-box classifiers are Neural Networks, SVMs, and ensemble classifiers, such as Random Forests, XGBoost or LightGBM. On the contrary, if a classifier is human-comprehensible, i.e., the reasons for its decisions are understandable by a human, we call it an *interpretable surrogate* classifier. Examples of such predictors include rule-based classifiers, decision trees, and decision sets [14, 2].

3.3 Black-box Global Explanations

In the landscape of XAI, two primary categories of explainers exist: local explainers, which focus on analysing a specific record, or global explainers, which aim to explain the overall behavior of a black-box model, considering explanations for all of the possible classes. In this paper we exploit TREPAN [10] as global explainer as well as a decision tree. TREPAN generates the surrogate explanation model by training a decision tree on an enriched version of the original dataset, with labels obtained through queries to the black-box model. TREPAN approaches the task of explaining neural networks as an inductive learning problem, simplified by the ability to query the black-box model during the process. This capability is a key aspect of the method, as it allows for obtaining record labels, selecting internal node splits, and determining whether a node exclusively covers records of the same class, which is a crucial parameter when the decision tree are used for explanation purposes.

3.4 Black-box Local Explanations

For this work we focus on post-hoc local explanation methods for tabular data, which exploit a surrogate model. In particular, among the different possibilities offered by the state-of-the-art, we have chosen LORE [17], an explanation method that outputs rules to explain the reasons that lead the black-box model to its final prediction. It also provides counterfactual rules that explain what changes are needed to obtain the opposite prediction. The reason behind the selection of this method is the fact that its rules closely resemble human reasoning, and the availability of both rules and counterfactual rules allows for an in-depth analysis of the neighborhood around the point under analysis.

LORE is a post-hoc, local and agnostic explanation method capable of explaining any type of black-box model, provided the ability to query the black-box for predictions. For this work, we utilized the latest version of the method [17], which is more stable compared to state-of-the-art methodologies and ensures a high fidelity of the surrogate model. Given a black-box model, denoted as b, performing a classification task, and a record x for which b predicted a target $\hat{y} = b(x)$, LORE generates a set of synthetic neighbors, denoted as Z, around x, by employing a genetic algorithm. This step aims to create a set of points that are close to the one we aim at explaining. Subsequently, LORE queries b to obtain the predicted labels for all synthetic records, resulting in $Z_y = b(Z)$. The synthetic dataset obtained is then used to train a surrogate decision tree model. The surrogate model is a simple, yet effective classifier in the vicinity of the point we need to explain. From this surrogate model, LORE extracts rules and counterfactual rules. The process of creating synthetic data, labeling them using b, and constructing a surrogate decision tree can be further customized. Specifically, based on the empirical results presented in the main paper, we opted to create multiple synthetic datasets to obtain a better description of the space around the point x. For each synthetic dataset generated, LORE trains a surrogate decision tree and then combines all of them to achieve greater stability.

3.5 Membership Inference Attack

In the paper [38], the authors assume that a machine learning algorithm is used to train a classifier b that captures the relationship between data records and their labels. In order to attack b trained on D_b^{train}, MIA defines an attack model $A(\cdot)$: it is a machine learning model able to discern if a record was part of the training dataset D_b^{train} or not. Note that, D_b^{train} is composed by $(x^i, y_o^i)_b$, where y_o^i is the true labels associated to x_b^i. In practice, the attack

$A(\cdot)$ is a binary classifier that predicts IN if the record was part of the training set or OUT otherwise. $A(\cdot)$ is trained on a dataset D_a^{train}: $(x^i, y^i)_a$, where each x_a^i is composed by the label predicted by the classifier b for a record under analysis and its probability vector $\overline{y^i}$ of length L obtained by querying a shadow model $s^i(\cdot)$ mimicking b; while y_a^i is the correct membership label and that can be IN or OUT. The attack model $A(\cdot)$ is a voting model composed of L machine learning models: one for each output class of the classifier model under attack. The key factor in this attack is the knowledge of the probability vector: given how the probabilities in $\overline{y}_b$ are distributed around the true value of the record, the attack model computes the membership probability $\Pr\{(x, y) \in D_b^{train}\}$, which is the probability that x belongs to the IN class, i.e. it is part of the training set. To obtain the dataset $(x^i, y^i)_a$, on which the MIA model $A(\cdot)$ is trained, the authors used *shadow models*. In the original paper the authors assume a black-box setting, in which there is no knowledge about either the type of classifier to be attacked or the training dataset used to train it. In the following we use the term black-box model to indicate the classifier to be attacked. To overcome the limitation of absence of knowledge on data and model, they employed a set of k shadow models $s^i(\cdot)$: machine learning models trained to mimic the decisions of the black-box model $b(\cdot)$ we would like to attack. These shadow models are trained on D_s^{train}: $(x^i, y^i)_s$, in which x_s^i has the same format and similar distribution w.r.t. to the dataset employed to train the black-box model X, while y_s^i is the predicted class obtained querying the black-box model $b(\cdot)$. After the training, we know which record was part of the training dataset (class IN) for each shadow model and which was part of the test one (class OUT). Hence, we can exploit this information to create a supervised training dataset for training the attack model $A(\cdot)$, which is D_a^{train}.

We highlight that the datasets employed for training the shadow models are disjoint from the unknown dataset used to train the black-box model. In [38] the authors tested different kinds of training data for the shadow models: (i) a *random* dataset, where data are randomly generated and then labelled querying the black-box model; (ii) a *statistical* dataset, in which the attacker knows the statistical distribution of the original training dataset. Hence, he/she can exploit this information to create a synthetic dataset; (iii) a *noise* dataset, in which the attacker knows a portion of data from the same distribution of the original training dataset but with some noise. These different types of training datasets for the shadow models allow for privacy attacks of different strengths: from the least severe attack, the random one, to the most powerful, i.e., the noise one.

3.6 Label Only Membership Inference Attack

A variant of MIA was designed in [8], which relaxes some requirements of the original attack. Given a black-box model b, LABELONLY $A_{LO}(\cdot)$ targets it by exploiting only the hard labels, i.e. the output predictions of the model under analysis. Hence, the probability vector $\overline{y^i}$, employed by MIA, is not exploited in LABELONLY. In particular, it develops a procedure that derives a model's robustness to perturbations and uses it as proxy for model confidence in its predictions. The basic intuition is that records which exhibit high robustness belong to the training dataset. $A_{LO}(\cdot)$ exploits a dataset D_s^{train} for training only one shadow model $s(\cdot)$, i.e., a ML model mimicking the decision of black-box model b. The dataset D_s^{train}: $(x^i, y^i)_s$ is composed of records with the same format and similar distribution w.r.t. to the dataset employed to train the black-box model b, and is labelled by the predicted class obtained querying b. After training the shadow model, we know which record was part of the training dataset (class IN) of the shadow model and which was part of the test one (class OUT). For each tuple x_s^i the algorithm generates a set of records resulting from its perturbation and labels the generated records using the trained shadow model. Analyzing the percentage of generated records having the same predicted class of x_s^i, the algorithm computes the robustness score of the black-box with respect to the x_s^i classification. The attack identifies a threshold by iteratively analyzing robustness scores assigned to records in the training and testing datasets of the shadow model. This threshold separates records into two classes: IN (training set of the shadow) and OUT (test set of the shadow). The attack then uses this threshold to classify new records as being part of the training set of the black-box model or not.

3.7 Agnostic Label Only Membership Inference Attack

The Agnostic Label Only Attack (ALOA), a variant of the LABELONLY attack, has been proposed recently [23]. Similar to the LABELONLY, ALOA does not require the access to the probability vector. However, this privacy attack has weaker assumptions with respect to the LABELONLY, since it does not need to know any kind of statistics about the data used for training the ML model we aim at attacking. In practice, the only requirement is to know the total number of variables to be able to query the black-box, with no information needed about the minimum, maximum, mean, or standard deviation.

Techinically, $A_{\text{aloa}}(\cdot)$ exploits a dataset D_s^{train} for training only one shadow model $s(\cdot)$, i.e., a ML model mimicking the decision of black-box model b. The dataset D_s^{train}: $(x^i, y^i)_s$ is composed of randomly generated records with the same format of the training dataset of the black-box model b, and is labelled by the predicted class obtained querying b. At this point, similarly to the other attacks, we know which record was part of the training dataset (class IN) of the shadow model and which was part of the test (class OUT). At this point, ALOA generates a set of synthetic records by perturbing the record under analysis. This perturbation procedure is completely agnostic and do not exploits any kind of statistics about the original dataset. Similarly as in the LABELONLY attack, the percentage of generated records having the same predicted class of the record under analysis is exploited to compute the robustness score of the black-box. At this point the robustness score is exploited to find the threshold that best separates the classes IN and OUT.

4 Framework for the Privacy Risk evaluation

In this section, we present the structure and components of our framework, REVEAL (pRivacy risk EValuation of Exposing surrogAte expLainers), which is designed to evaluate the privacy exposure of black-box models and assess how this exposure may change when they are explained using local or global surrogate-based explainers.

This framework, depicted in Figure 1, consists of three main modules: *Attack-Training*, which trains the attack models to be simulated, *Attack-Application*, which executes the trained attacks and *Attack-Evaluation*, which quantifies the privacy exposure introduced by an explainer. Algorithm 1 reports the pseudo-code of the whole assessment framework.

The instantiation of the three modules depends on the assumed threat model, the type of privacy attack to be performed, and the type of surrogate explainer used to explain the black-box model. We emphasize that our framework is specifically designed for surrogate-based explainers, i.e. explainers that rely on a surrogate model to generate explanations. However, our approach is not limited to any particular type of data: whether the input to the black-box model, and consequently the explainer, is tabular, image, or time series data,

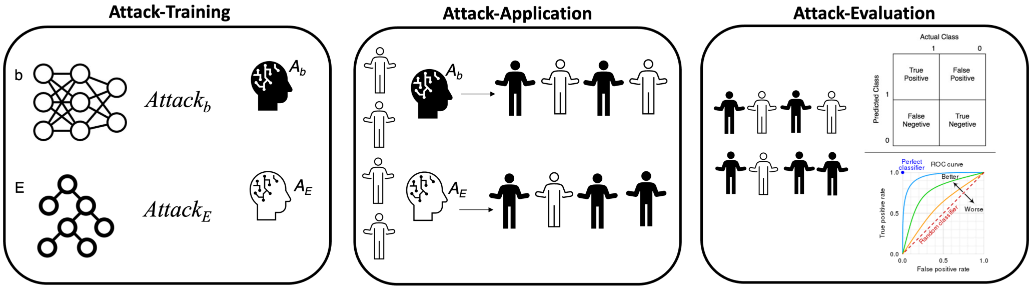

Figure 1: Schema of REVEAL, for Privacy Exposure of black-box models and their explainers. The framework is composed by three modules: the first one, *Attack Training* is devoted to train the chosen attack against the black-box and its explainer. Following, the *Attack Application* applies the trained privacy attacks to a dataset predicting the membership of each record. Lastly, the *Attack Evaluation* evaluates the changes of the privacy exposure when attacking the black-box and its explainer. As an example, in the context of global explanations, we consider a Random Forest as the black-box model (b) and TREPAN as the explainer (E). At this point, we train two separate attack models: Attack$_b$ for the RF and Attack$_E$ for TREPAN. Once trained, we perform the actual attacks on RF and TREPAN and evaluate the performance of these attacks against both models. The same setting can be considered for the local explanation case, with the difference that, instead of TREPAN, we can use methods such as LORE, which provide explanations for individual records. In this case, the final attack may involve an ensemble of the various local attacks created.

REVEAL remains unaffected. In the following, we describe the objective and role of each module within the framework.

Algorithm 1: PrivacyRiskExposure(b, E, D^{test}, $Attack_b$, $Attack_E$, BK)

```
$(A_b, A_E) \leftarrow$ Attack-Training$(b, E, Attack_b, Attack_E, BK)$ ;
$(D^{\text{test}}_{\text{b-member}}, D^{\text{test}}_{\text{E-member}}) \leftarrow$ Attack-Application$(A_b, A_E, D^{\text{test}})$ ;
$[\Delta_{Acc}, \Delta_P, \Delta_R] \leftarrow$ Attack-Evaluation$(D^{\text{test}}_{\text{b-member}}, D^{\text{test}}_{\text{E-member}})$ ;
return $[\Delta_{Acc}, \Delta_P, \Delta_R, \Delta_{F_1}]$
```

Attack-Training Module Given the back-box model b and its explainer E, the first module aims at learning two privacy attack models: the first one, namely A_b, is tailored to attack the black-box model b, while the second one, referred to as A_E, is tailored to attack the explainer E. As explained in Section 2, different attacks can be conducted for auditing a machine learning model. However, one of the most used attacks is the membership inference attack [38], aiming at inferring the membership of records to the training data of the machine learning model. This type of attack is also the foundation of other attacks aiming at extracting records from training data [5]. In this work, we propose to instantiate the model under analysis with learning algorithms for training these kinds of attacks. We highlight that the function $Attack_b(\cdot)$, aiming at learning the attack model A_b, can be different from $Attack_E(\cdot)$ that is used for learning A_E. This difference could be due to the fact that the black-box and the explainers might be ML models completely different that do not allow the attack under similar assumptions. For example, we could have a black-box that does not return the confidence vector for each prediction while its explainer could return it. Consequently, $Attack_E(\cdot)$ could exploit this additional information. The

Algorithm 2: Attack-Training $(b, E, Attack_b, Attack_E, BK)$

```
1 $X_a^{\text{train}} \leftarrow GenerateAttackDataset(BK)$ ;
2 $Y_b \leftarrow b(X_a^{\text{train}})$ ;
3 $A_b \leftarrow$ Learning: $Attack_b(X_a^{\text{train}}, Y_b)$ ;
4 $Y_E \leftarrow E(X_a^{\text{train}})$ ;
5 $A_E \leftarrow$ Learning: $Attack_E(X_a^{\text{train}}, Y_E)$ ;
6 return $A_b,\ A_E$
```

two functions executed in this module may be implemented using one of the algorithms available in the literature, such as the Membership Inference Attack [38], the Label-Only Attack [8], the ALOA Attack [23], all introduced in Section 3, or any other attack for ML models. Moreover, global and local explainers might be required to design and develop a slightly different learning procedure for the attack. As an example, in the following, we propose learning an ensemble of attack models for attacking local explainers and assessing the privacy risks introduced by these types of explainers. The pseudo-code of this module is reported in Algorithm 2. We highlight that before training the two attacks, this module also generates the dataset X_a^{train} useful for learning the attacks. Such a dataset is labeled by using both the black-box (line 2, Alg. 2) and the explainer (line 4, Alg. 2). The type of attack dataset generated strongly depends on the background knowledge of the adversary BK. For example, if an adversary knows the distribution of the black-box training data, the attack can exploit this knowledge to generate the attack dataset. The performance of the attacks can be heavily affected by the properties of this dataset.

Algorithm 3: Attack-Application$(A_b, A_E,\ D^{\text{test}})$

```
1 $D^{\text{test}}_{\text{b-member}} \leftarrow A_b(D^{\text{test}})$ ;
2 $D^{\text{test}}_{\text{E-member}} \leftarrow A_E(D^{\text{test}})$ ;
3 return $(D^{test}_{b\text{-}member}, D^{test}_{E\text{-}member})$
```

Attack-Application Module The second module of our framework is called *Attack- Application* and applies the attack models learned in the previous module A_b and A_E for inferring the membership of individual records to the training of b. The pseudo-code of this module is reported in Algorithm 3. In particular, given a set of records D^{test}, this module conducts the two attacks against the black-box and the explainer, respectively, and for each record outputs their membership prediction inferred by the two attack models, i.e., the labelled datasets $D^{\text{test}}_{\text{b-member}}$ and $D^{\text{test}}_{\text{E-member}}$ (line 1-2, Alg. 3). The two sets of labelled records are the base for computing and assessing the *Privacy Risk Exposure* for both the black-box model and its explainer. The instantiation of this module strongly depends on the attacks learned in the previous module and the type of explainers (global vs. local). Indeed, later in this chapter, we will show that this module is the main difference between the assessment of global and local explainers.

Attack-Evaluation Module The output of the second module is then fed into the third and final module, the *Attack-Evaluation* module. This module aims to analyze and quantify the change of privacy risk exposure between the black-box model b and its explainer E.

```
Algorithm 4: Attack-Evaluation($D^{\text{test}}_{\text{b-member}}$, $D^{\text{test}}_{\text{E-member}}$)
1 $C_{\text{b-member}} \leftarrow ConfusionMatrix(D^{\text{test}}_{\text{b-member}})$ ;
2 $C_{\text{E-member}} \leftarrow ConfusionMatrix(D^{\text{test}}_{\text{E-member}})$ ;
3 $[\Delta_{Acc}, \Delta_P, \Delta_R, \Delta_{F_1}] \leftarrow Compute\Delta(C_{\text{b-member}}, C_{\text{E-member}})$ ;
4 return $[\Delta_{Acc}, \Delta_P, \Delta_R, \Delta_{F_1}]$
```

The analysis can be performed using different metrics that evaluate the performance of the attack models in predicting the membership of the individual records to the training data of b. This module first evaluates the confusion matrix for the attack against the black-box, A_b (line 1, Algorithm 4), then for the attack against the explainer, A_E (line 12 Algorithm 4). From these partial results, the module performs an overall evaluation in terms of standard ML metrics. In particular, this module computes the difference in privacy exposure in terms of accuracy (Δ_{Acc}), precision (Δ_P), recall (Δ_R) and f-measure (Δ_{F_1}) of the two attack models. In other words, each Δ_μ is computed as $\Delta_\mu = \Delta^E_\mu - \Delta^b_\mu$, where μ denotes one of the metrics among accuracy, precision and recall. Analyzing only accuracy for evaluating membership inference attacks could be inadequate because these metrics associate equal costs to false positives (false memberships) and false negatives (false non-memberships). The first type of error reduces the utility of the attack, while the second one reduces the identification of real members. An attack should maximize the true positive rate (or recall) because it measures how many members are identified. We highlight that negative values of Δ for a given measure μ mean that the explainer tends to mitigate the privacy risks of the black-box, i.e., the explanation procedure is confusing the attack; positive values of Δ instead highlight higher privacy risks due to the level of transparency introduced by the explainer; lastly, $\Delta = 0$ means that pairing an explainer with a black-box classifier is not increasing the privacy risks.

4.1 Instantiation of REVEAL for Global and Local Explainers

REVEAL is a framework for assessing the privacy exposure in black-boxes and their explainers. The framework presented is generic and can work with any ML model, as well as any surrogate-based explainer, but needs to be instantiated differently depending on the attack considered due to the different background knowledge possessed by the adversary. In this work, we propose to instantiate such a framework with attacks belonging to the family of *membership* attacks and we investigate the impact of different levels of adversary background knowledge on the success of the attack. In particular, we investigate the privacy exposure of global and local explainers under the attacks MIA, LABELONLY and ALOA, described in Section 3.

4.1.1 REVEAL for Global Explainers

Instancing REVEAL for the assessment of the privacy risk exposure of global explainers is straightforward. A global explainer based on a surrogate model E is a ML model that imitates the global behaviour of a black-box classifier b. As a consequence, it is enough to follow the procedure described in the previous section, implementing the training of one of the membership-based attack models presented above. In particular, in the first module, *Attack-Training*, trains both *(i)* a privacy attack, named A_b, against b is trained, being it MIA, LABELONLY or ALOA; and *(ii)* a privacy attack, named A_E, against the explainer E. Then, these two attacks are fed into the *Attack-Application* module, which applies these attacks to a test dataset, namely D^{test}. The result will be to obtain two labelled datasets: one which for each element of the dataset has the membership class IN or OUT determined by the attack A_b, and one with the class determined by the attack A_E. Lastly, the *Attack-Evaluation* module quantifies the probability of success of the two attacks, against the black-box and the explainer, computing the difference in the performance of both concerning *precision*, *recall*, and *accuracy*.

4.1.2 REVEAL for Local Explainers

When employing the REVEAL framework to assess the privacy vulnerability of a black-box model and its local models, it is essential to tailor the attack methodology to the specific scenario being analyzed, where E represents a collection of local surrogate models. In this context, each local explainer is customized to describe a small portion of the decision boundary of the black-box. Therefore, to ensure that the entire decision boundary of the black-box model b is properly described, it is imperative to consider a variety of local explainers that capture different types of local knowledge. This means that if an adversary wants to jeopardize the privacy of a black-box attacking its local explainer, it needs to generate a set of local explainers that all together approximate the black-box's global behavior. To this end, we propose a privacy attack procedure designed to target local surrogate-based explainers. Specifically, the procedure assumes E as a set of local surrogates, i.e., $E = e_1, e_2, \dots, e_n$. Following the pseudo-code outlined earlier in Algorithm 2, the $Attack_E$ is computed as an ensemble of multiple attacks, with one attack tailored for each local surrogate model in E. The resulting ensemble of attacks is denoted as $A_E = Ae_1, A_{e_2}, \dots, A_{e_n}$ and is passed, along with the attack tailored for the black-box A_b, to the *Attack-Application* module. In this setting, the module needs to evaluate the effectiveness of the ensemble of attacks A_E against the attack for the black-box A_b. The application of A_E can be instantiated in different ways, depending on the specific information assumed by the attack, e.g. the different background knowledge the attacker may have. In the following, we present two ways for implementing *Attack-Application* in the local setting depending on the knowledge the attack produces. In particular, we consider two approaches: the *Confidence Vector* Approach, based on the prediction probabilities vectors, applicable to every membership attack based on the creation of ML attack models, such as the original MIA; and the *Threshold* Approach, tailored for the attacks which do not create a ML model, but a thresholding procedure, such as LABELONLY and ALOA.

For the **Confidence Vector Approach**, we apply an evaluation procedure that exploits the prediction probabilities vectors outputted by the attack models. This setting is tailored for methods such as MIA, which trains a ML attack model for each target output from the black-box model. Having created these attacks based on ML models, we assume to have access to the prediction confidence vectors, $c = [c_{\text{IN}}, c_{\text{OUT}}]$, where c_{IN} is the probability that the record belongs to class IN, while c_{OUT} is the probability that the record belongs to class OUT and the sum of all the two elements is equal to 1. Hence, we exploit this information to identify among the different attacks only the ones that are the most confident record-wise. Technically, for each record x, we apply all the attack models, obtaining a confidence vector for each one, i.e., $C_x = \{c^{A_1}, c^{A_2}, \dots, c^{A_n}\}$, where n is the number of attacks for the n local explainers. At this point, for each vector c^{A_i}, we compute the absolute difference between the two probabilities, i.e., $d_i = |c_{\text{IN}}^{A_{e_i}} - c_{\text{OUT}}^{A_{e_i}}|$.

Once we get the corresponding d value for each attack model, we select only the attack models expressing significant confidence in their decisions. To this end, we select the models A_{e_j} having a d_j value above the average. In particular, we use the following constraint for selecting the attack set: $\{A_{e_j}|d_j \geq (avg(d_1, d_2, \ldots, d_n) + \sigma(d_1, d_2, \ldots, d_n))\}$, where σ is the standard deviation. Among the top attack models selected, we apply a majority voting procedure to select the final membership prediction for each record.

In the case of **Threshold Approach**, the *Attack-Evaluation* strategy is tailored for membership attacks that do not train ML models as attacks but use a thresholding procedure. Examples of attacks of this family are LABELONLY and ALOA. In this setting, we exploit the different information available, which is the threshold found and used by each attack for the membership prediction. Given the record x under analysis, by applying the attack A_{e_i} and we obtain a robustness score $s^{A_{e_i}}$, which is compared to the score threshold $st^{A_{e_i}}$ for determining IN or OUT class. Hence, we exploit the absolute distance between the robustness score and the score threshold (i.e., $d_i = |s^{A_{e_i}} - st^{A_{e_i}}|$) to identify the most reliable attacks. In particular, we are interested in the attacks which have a greater distance between the robustness score of the record and the score threshold. We select only the top attack models, exploiting the *elbow* method, i.e., we select the most important models with a d value greater or equal to the one corresponding to the knee in the curve of the ordered d values Formally, we select the following set of attack models: $\{A_{e_j}|d_j > \text{elbow}(d_1, \ldots, d_n)\}$. We apply a majority voting strategy to obtain the final membership prediction on the set of attack models selected. These two evaluation methods presented are only two possible initializations, dependent on the privacy attack considered.

5 Experiments

In this section, we instantiate our framework to conduct a series of experiments, exploiting various datasets, black-box models, and explainers. Our primary objective is to analyze the behavior of black-box and explainer systems, specifically focusing on how privacy exposure may vary across different experimental settings. We present the datasets and the ML models employed in Section 5.1. Following, we present the application of REVEAL, exploiting the MIA, LABELONLY and ALOA, in Section 5.1.

5.1 Data, Machine Learning models and Explainers

For validating the framework presented in this paper we employed three tabular datasets, each with particular characteristics. Although REVEAL is data-agnostic, we focused our analysis on tabular data due to the availability of privacy attacks targeting both ML models and surrogate-based explainers in this context.

We select ADULT, a benchmarking dataset composed of $48,842$ records and 15 variables, both numerical and categorical. It describes employees with information like age, job, capital loss, capital gain, marital status, etc. The labels have values $<= 50K$ or $> 50K$ (in the following referred to as respectively as Class 0 or Class 1), indicating whether the person will earn more or less than $50k$ in a fiscal year. We chose this dataset on the basis that it is often used for benchmarking and has also been used in the papers of MIA, ALOA and LABELONLY, which we exploit in this work. This dataset was also used as a validation set of the attack for the MIA and LABELONLY. We also consider BANK, which is a public dataset containing information of the customers of a bank, with the objective of classifying the people in good or bad creditors. It is formed by $150,000$ records and 10 numerical variables, with information like age, monthly income and the number of loans already opened. The selection of this dataset is due to the huge amount of records available as well as the peculiarity of having only numerical variables. Lastly, we also consider the SYNTH dataset, which is a synthetic dataset generated by exploiting a Gaussian mixture model. It has $30,000$ records and 30 numerical variables, with 3 classes. The selection of this dataset is due to the multi-classification problem and to test the behaviour of the attack in a controlled environment due to the synthetic creation of the dataset.

Regarding the pre-processing, for ADULT, we removed the null values and analyzed the Pearson correlation among the variables, dropping some of them to obtain a correlation degree less than 80%. For the remaining categorical variables, we applied a one-hot encoding. For BANK, we removed the null values, and the correlation analysis did not highlight any correlation value higher than 75%. Thus, we did not drop any variable. No further pre-processing was needed since the variables were all numerical. For SYNTH, instead, we did not perform any kind of pre-processing since the dataset was synthetically generated.

After the pre-processing step, we split each dataset into two subsets: (i) 70% of the original dataset (called D_b) is used to train and test the black-box models; (ii) the remaining 30% of the pre-processed data dataset (called D_s) is used for the learning process of the different attacks, in particular for those attacks that require a minimum background knowledge information about the original data distribution durign the creation of the shadow models.

Table 1 Predictive performance of the models for ADULT, SYNTH and BANK dataset on the test set. The results are validated with 3-fold cross-validation (we provide the mean and the standard deviation between brackets). This table highlights the extremely good predictive performance of TREPAN w.r.t. DT and RF, which is almost always the best model, except for SYNTH. TREPAN was trained to exploit an enriched dataset, but in this case we tested the predictive performance on the same test set of the black-boxes for comparison purposes.

Data	Balance	Metric	DT	RF	TREPAN-RF	NN	TREPAN-NN
ADULT	$C_1 = 24\%$ $C_0 = 76\%$	F_{1_1}	0.63 (0.02)	0.70 (0.02)	**0.98** (0.00)	0.67 (0.02)	**0.77** (0.02)
		P_1	0.60 (0.01)	0.69 (0.02)	**0.99** (0.00)	0.69 (0.02)	**0.82** (0.00)
		R_1	0.58 (0.05)	0.87 (0.03)	**0.98** (0.01)	0.67 (0.03)	**0.73** (0.01)
		F_{1_0}	0.90 (0.00)	0.86 (0.00)	**0.99** (0.00)	0.89 (0.00)	**0.99** (0.00)
		P_0	0.87 (0.01)	0.95 (0.00)	**0.98** (0.00)	0.90 (0.00)	**0.99** (0.00)
		R_0	0.92 (0.01)	0.80 (0.01)	**0.99** (0.00)	0.89 (0.01)	**0.98** (0.01)
	Balance	**Metric**	**DT**	**RF**	**TREPAN-RF**	**NN**	**TREPAN-NN**
SYNTH	$C_2 = 33\%$	F_{1_2}	0.77 (0.01)	0.99 (0.01)	0.95 (0.01)	1.00 (0.00)	**0.72** (0.01)
		P_2	0.96 (0.01)	0.98 (0.01)	**0.94** (0.00)	1.00 (0.00)	**0.75** (0.00)
		R_2	0.96 (0.01)	1.00 (0.00)	**0.98** (0.01)	0.99 (0.01)	**0.70** (0.01)
	$C_1 = 33\%$	F_{1_1}	0.81 (0.02)	0.89 (0.01)	**0.82** (0.01)	**0.93** (0.01)	0.67 (0.01)
		P_1	0.83 (0.00)	0.88 (0.00)	**0.84** (0.00)	**0.94** (0.00)	0.64 (0.00)
		R_1	0.80 (0.00)	0.89 (0.01)	**0.80** (0.01)	**0.93** (0.01)	0.72 (0.01)
	$C_0 = 33\%$	F_{1_0}	0.80 (0.02)	0.88 (0.01)	**0.82** (0.01)	**0.93** (0.01)	0.89 (0.01)
		P_0	0.80 (0.00)	0.89 (0.00)	**0.80** (0.00)	**0.92** (0.00)	0.90 (0.00)
		R_0	0.82 (0.00)	0.88 (0.01)	**0.86** (0.01)	**0.94** (0.01)	0.88 (0.02)
	Balance	**Metric**	**DT**	**RF**	**TREPAN-RF**	**NN**	**TREPAN-NN**
BANK	$C_1 = 8\%$ $C_0 = 92\%$	F_{1_1}	0.35 (0.01)	0.77 (0.01)	**0.99** (0.01)	0.78 (0.01)	**0.84** (0.01)
		P_1	0.38 (0.01)	0.83 (0.01)	**0.98** (0.02)	0.77 (0.01)	**0.86** (0.00)
		R_1	0.34 (0.01)	0.75 (0.04)	**0.99** (0.01)	0.76 (0.04)	**0.82** (0.01)
		F_{1_0}	0.95 (0.02)	0.92 (0.01)	**0.99** (0.01)	0.77 (0.01)	**0.95** (0.01)
		P_0	0.95 (0.00)	0.91 (0.00)	**0.99** (0.00)	0.78 (0.00)	**0.96** (0.02)
		R_0	0.95 (0.00)	0.92 (0.01)	**0.98** (0.01)	0.79 (0.01)	**0.95** (0.01)

On each of the datasets selected we train different ML models: a *Decision Tree*, in the following referred to as DT, a simple, explainable by design method, which is exploited as a benchmarking; a *Random Forest* (RF), an ensemble method based on trees, able to achieve extremely good prediction performance with tabular data; and a feedforward *Neural Network* (NN). We chose these ML methods to examine the behaviour of our framework on models that have completely different structural characteristics. The results of the training of these ML models are reported in Table 1. Overall, the performance of the models is good, with the exception of the DT, which suffers greatly from the imbalance between the classes, especially high in the case of the BANK dataset. However, this result is in line with the state-of-the-art, which shows that DT suffers greatly from noise and imbalanced data.

After training the ML models, we also train the respective explainers. Given the inherent interpretability of DT, we do not need to explain them using a XAI method: we will only exploit the DT to compare the results obtained from the application of other methods. For the *global* case, we consider TREPAN, a tree-based explainer fitted on an enhanced version of the original training dataset, labelled by the black-box model b. Therefore, we train a TREPAN-RF model for explaining the RF, and a TREPAN-NN model for the NN. The performance of the TREPAN model is reported in Table 1, from which it is possible to see that the performance of the TREPAN models is extremely good for all the datasets. For the *local* case, we select LORE, a post-hoc agnostic explainer that exploits a local DT surrogate model to extract rules and counterfactual rules. In this case, we train one local surrogate model for each record to explain. The average fidelity of these models is 0.97 ± 0.08.

5.2 Reveal evaluation

After training the black-box models and their explainers, we can now test the performance of REVEAL. Following the experimental setting presented in [38], we consider two settings for the fitting of the shadow models: the worst case scenario, called *noise* dataset from now on, and the best case scenario, called *random* dataset. In the case of the noise dataset, we assume the attacker has access to a noise version of a set of data from the same distribution of the data exploited in the training set of the black-box. Technically, we add 10% of noise to a piece of original dataset not exploited during the training of the black-box. For the random case, instead, we assume the attacker has no knowledge about the dataset used for training the black-box, apart of the number of variables of the original data. Therefore, the attacker randomly generates and labels a dataset by querying the black-box. The choice of these two types of datasets is due to the different settings they create: the noise dataset assumes a favourable setting for the attacker, who, through some public information or misappropriation of information, can obtain a piece of data from the same distribution as the original one, albeit with some noise. This setting is unrealistic, but it is also where the MIA allows greater privacy exposure. In addition, LABELONLY requires knowledge of statistical information from the original data, so this setting is in line with the assumptions of this attack. In the second setting, on the other hand, the attacker has no knowledge of the data, and it is, therefore, a more realistic setting. At the same time, it is also relevant to assess the performance of the attacks in this context, as having a privacy risk in this case is much worse than the previous setting because it requires less knowledge on the part of the attacker. In addition, it is also interesting to analyze the behaviour of the various attacks in the *random* setting: based on the work in the literature, in this setting we expect to have a decrease in privacy exposure for MIA and LABELONLY, while performance should remain roughly similar for the ALOA case, which is specifically developed for this setting.

For each combination of black-box model, explainer and kind of dataset to generate the shadow models (e.g. random or noise), we train MIA, LABELONLY and ALOA, both for the global and the local explanations. Due to the different methodologies applied for the local and global case, in the following we present the results separately.

5.3 Results attacking the Global Explainers

Table 2 Results of the application of MIA, LABELONLY and ALOA with the setting *noise* for training the shadow models, attacking global explainers. For each attack is reported the P (Precision) and R (Recall) and F_1 score (harmonic mean of P and R) for the IN class, which is the class of the records correctly identified by the attacker. The values reported are the mean over a 3 fold cross validation, with the standard deviation between brackets.

-		MIA			LABELONLY			ALOA		
Dataset	**Metric**	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF
ADULT	$F_{1_{\text{In}}}$	**0.79** (0.01)	0.70 (0.01)	**0.77** (0.01)	**0.73** (0.01)	0.81 (0.01)	**0.79** (0.00)	**0.78** (0.01)	0.81 (0.01)	**0.79** (0.00)
	P_{In}	**0.80** (0.02)	0.80 (0.03)	**0.80** (0.00)	0.81 (0.01)	**0.82** (0.00)	0.80 (0.00)	0.81 (0.01)	**0.79** (0.00)	0.79 (0.00)
	R_{In}	**0.77** (0.01)	0.67 (0.01)	**0.72** (0.02)	**0.70** (0.02)	0.81 (0.01)	**0.81** (0.03)	**0.76** (0.02)	0.80 (0.01)	**0.81** (0.03)
	Δ_R	**0.10**	-	**0.05**	**-0.09**	-	**0.00**	**-0.04**	-	**0.01**
	Metric	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	**0.79** (0.01)	0.63 (0.02)	**0.70** (0.01)	**0.73** (0.01)	0.73 (0.02)	**0.79** (0.01)	**0.78** (0.01)	0.64 (0.02)	**0.78** (0.01)
	P_{In}	**0.80** (0.02)	0.79 (0.00)	**0.79** (0.03)	0.81 (0.01)	**0.80** (0.00)	**0.79** (0.03)	0.81 (0.01)	**0.81** (0.00)	**0.78** (0.03)
	R_{In}	**0.77** (0.01)	0.53 (0.03)	**0.64** (0.00)	**0.70** (0.02)	0.67 (0.03)	**0.80** (0.00)	**0.76** (0.02)	0.53 (0.03)	**0.79** (0.00)
	Δ_R	**0.24**	-	**0.11**	**0.03**	-	**0.13**	**0.23**	-	**0.26**
SYNTH	**Metric**	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF
	$F_{1_{\text{In}}}$	**0.77** (0.01)	0.76 (0.01)	**0.78** (0.00)	**0.85** (0.01)	0.98 (0.01)	**0.97** (0.00)	**0.85** (0.01)	0.72 (0.01)	**0.83** (0.00)
	P_{In}	0.70 (0.01)	**0.70** (0.00)	0.70 (0.00)	0.86 (0.01)	**0.83** (0.00)	0.82 (0.00)	0.86 (0.01)	**0.84** (0.00)	0.72 (0.00)
	R_{In}	**0.85** (0.02)	0.82 (0.01)	**0.87** (0.03)	**0.84** (0.02)	0.82 (0.01)	**0.93** (0.03)	**0.84** (0.02)	0.62 (0.01)	**0.80** (0.03)
	Δ_R	**0.03**	-	**0.05**	**0.02**	-	**0.11**	**0.22**	-	**0.20**
	Metric	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	**0.77** (0.01)	0.78 (0.02)	**0.79** (0.01)	**0.85** (0.01)	0.86 (0.02)	**0.81** (0.01)	**0.85** (0.01)	0.85 (0.02)	**0.72** (0.01)
	P_{In}	0.70 (0.01)	**0.70** (0.00)	**0.79** (0.03)	0.86 (0.01)	**0.80** (0.00)	**0.81** (0.03)	0.86 (0.01)	**0.81** (0.00)	**0.75** (0.03)
	R_{In}	**0.85** (0.02)	0.88 (0.03)	**0.90** (0.00)	**0.84** (0.02)	0.90 (0.03)	**0.90** (0.00)	**0.84** (0.02)	0.90 (0.03)	**0.83** (0.00)
	Δ_R	**0.03**	-	**0.02**	**-0.06**	-	**0.00**	**-0.06**	-	**-0.07**
BANK	**Metric**	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF
	$F_{1_{\text{In}}}$	0.67 (0.01)	0.71 (0.03)	**0.75** (0.03)	**0.79** (0.01)	0.78 (0.01)	**0.79** (0.00)	**0.79** (0.01)	0.77 (0.01)	**0.77** (0.00)
	P_{In}	0.65 (0.02)	0.67 (0.02)	**0.67** (0.00)	0.80 (0.01)	**0.80** (0.00)	0.79 (0.00)	0.80 (0.01)	**0.65** (0.00)	0.79 (0.00)
	R_{In}	0.67 (0.01)	0.80 (0.02)	**0.85** (0.00)	**0.78** (0.02)	0.76 (0.01)	**0.80** (0.03)	**0.78** (0.02)	0.78 (0.01)	**0.79** (0.03)
	Δ_R	**-0.10**	-	**0.05**	**0.02**	-	**0.04**	**0.00**	-	**0.01**
	Metric	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	**0.67** (0.01)	0.30 (0.00)	**0.69** (0.00)	**0.78** (0.01)	0.78 (0.02)	**0.79** (0.01)	**0.79** (0.01)	0.79 (0.02)	**0.79** (0.01)
	P_{In}	0.65 (0.02)	**0.79** (0.01)	0.65 (0.00)	0.80 (0.01)	**0.80** (0.00)	**0.80** (0.03)	0.80 (0.01)	**0.80** (0.00)	**0.80** (0.03)
	R_{In}	**0.67** (0.01)	0.25 (0.02)	**0.72** (0.02)	**0.78** (0.02)	0.77 (0.03)	**0.78** (0.00)	**0.78** (0.02)	0.78 (0.03)	**0.80** (0.00)
	Δ_R	**0.42**	-	**0.47**	**0.01**	-	**0.01**	**0.00**	-	**0.02**

To evaluate REVEAL, we attack both the black-box models and their surrogate-based explainers employing three different attacks: MIA, LABELONLY and ALOA.

For training each MIA, we train 6 shadow models with the objective of mimicking the black-boxes. The shadow models are trained employing the best set of hyper parameters found using a grid search. All of the shadow models have an accuracy above 80%. Then, from the shadow models, we extract the supervised training dataset D_a^{train} to train the attack model. We remark that the MIA assumes the attack model as an ensemble model composed of a ML model for each label L. Hence, in our case, we obtain two (or three for the SYNTH dataset) RF attack models for each attack. Also in this case, for the different attack models, we first search for the best set of hyperparameters, obtaining an accuracy above 94% for all the models, when tested on a portion of test data D_a^{test}.

For the LABELONLY and ALOA, we have just one shadow model, a RF as for the MIA, with an accuracy above 80%. After fitting the shadow model, both models require the computation of the robustness score, creating 1000 perturbations for each record, and the final attack model is not a ML model but a thresholding model, adaptively selected depending

Table 3 Results of the application of MIA, LABELONLY and ALOA with the setting *random* for training the shadow models, attacking global explainers. For each attack is reported the F_1, Precision and Recall for the IN class, which is the class of the records correctly identified by the attacker. The values reported are the mean over a 3 fold cross validation, with the standard deviation between brackets.

-		MIA			LABELONLY			ALOA		
Dataset	**Metric**	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF	DT	RF	TREPAN-RF
ADULT	$F_{1_{\text{In}}}$	**0.72** (0.01)	0.49 (0.01)	**0.78** (0.01)	**0.30** (0.01)	0.46 (0.01)	**0.78** (0.01)	**0.77** (0.01)	0.82 (0.01)	**0.77** (0.01)
	P_{In}	**0.78** (0.02)	0.77 (0.01)	**0.79** (0.00)	**0.78** (0.02)	0.77 (0.01)	**0.79** (0.00)	**0.82** (0.02)	0.78 (0.01)	**0.77** (0.00)
	R_{In}	**0.66** (0.01)	0.36 (0.01)	**0.77** (0.00)	**0.55** (0.01)	0.35 (0.01)	**0.80** (0.00)	**0.76** (0.01)	0.78 (0.01)	**0.82** (0.00)
	Δ_R	**0.30**	-	**0.41**	**0.20**	-	**0.50**	**0.02**	-	**0.04**
	Metric	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	**0.72** (0.01)	0.43 (0.02)	**0.77** (0.01)	**0.30** (0.01)	0.33 (0.02)	**0.67** (0.01)	**0.77** (0.01)	0.63 (0.02)	**0.77** (0.01)
	P_{In}	**0.78** (0.02)	0.70 (0.01)	**0.78** (0.01)	**0.78** (0.02)	0.77 (0.01)	**0.68** (0.01)	**0.82** (0.02)	0.81 (0.01)	**0.77** (0.01)
	R_{In}	**0.66** (0.01)	0.32 (0.01)	**0.76** (0.00)	**0.55** (0.01)	0.52 (0.01)	**0.66** (0.00)	**0.76** (0.01)	0.52 (0.01)	**0.77** (0.02)
	Δ_R	**0.33**	-	**0.44**	**0.03**	-	**0.14**	**0.24**	-	**0.25**
SYNTH	**Metric**	DT	RF	TREPAN-RF	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	**0.80** (0.01)	0.70 (0.01)	**0.77** (0.00)	**0.79** (0.01)	0.80 (0.01)	**0.80** (0.00)	**0.86** (0.03)	0.71 (0.00)	**0.83** (0.00)
	P_{In}	0.71 (0.01)	**0.85** (0.00)	0.70 (0.00)	0.80 (0.01)	**0.78** (0.00)	0.79 (0.00)	0.84 (0.01)	**0.82** (0.00)	0.73 (0.00)
	R_{In}	**0.98** (0.04)	0.78 (0.01)	**0.84** (0.03)	**0.78** (0.04)	0.76 (0.01)	**0.80** (0.03)	**0.83** (0.03)	0.70 (0.00)	**0.80** (0.03)
	Δ_R	**0.20**	-	**0.06**	**0.02**	-	**0.04**	**0.13**	-	**0.10**
	Metric	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	**0.80** (0.01)	0.45 (0.02)	**0.80** (0.04)	**0.79** (0.01)	0.82 (0.02)	**0.73** (0.02)	**0.86** (0.03)	0.84 (0.01)	**0.72** (0.00)
	P_{In}	0.71 (0.01)	**0.69** (0.00)	**0.70** (0.04)	0.80 (0.01)	**0.80** (0.00)	**0.68** (0.01)	0.84 (0.01)	**0.80** (0.10)	**0.72** (0.02)
	R_{In}	**0.98** (0.04)	0.33 (0.02)	**0.90** (0.02)	**0.78** (0.04)	0.77 (0.02)	**0.77** (0.02)	**0.83** (0.03)	0.89 (0.02)	**0.86** (0.01)
	Δ_R	**0.65**	-	**0.57**	**0.01**	-	**0**	**0.06**	-	**0.03**
BANK	**Metric**	DT	RF	TREPAN-RF	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	0.70 (0.02)	0.68 (0.03)	**0.73** (0.03)	0.72 (0.02)	0.70 (0.03)	**0.72** (0.03)	0.76 (0.20)	0.85 (0.03)	**0.76** (0.01)
	P_{In}	0.65 (0.04)	0.71 (0.02)	**0.64** (0.06)	0.79 (0.03)	0.65 (0.02)	**0.76** (0.06)	0.77 (0.01)	0.64 (0.02)	**0.79** (0.01)
	R_{In}	0.70 (0.10)	0.65 (0.02)	**0.85** (0.10)	0.76 (0.12)	0.73 (0.02)	**0.75** (0.10)	0.75 (0.02)	0.78 (0.02)	**0.80** (0.00)
	Δ_R	**0.05**	-	**0.20**	**0.03**	-	**0.02**	**0.03**	-	**0.02**
	Metric	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN	DT	NN	TREPAN-NN
	$F_{1_{\text{In}}}$	0.70 (0.02)	0.27 (0.04)	**0.65** (0.02)	0.72 (0.02)	0.47 (0.04)	**0.57** (0.01)	0.76 (0.20)	0.77 (0.00)	**0.79** (0.01)
	P_{In}	0.65 (0.04)	**0.70** (0.10)	0.65 (0.03)	0.79 (0.03)	**0.65** (0.06)	0.66 (0.00)	0.77 (0.01)	**0.10** (0.10)	0.78 (0.00)
	R_{In}	0.70 (0.10)	0.23 (0.02)	**0.69** (0.10)	0.76 (0.12)	0.46 (0.00)	**0.61** (0.01)	0.75 (0.02)	0.74 (0.11)	**0.78** (0.09)
	Δ_R	**0.47**	-	**0.46**	**0.30**	-	**0.20**	**0.01**	-	**0.04**

on the data in input.

Regarding the *global* explainers, the results are reported in Table 2 and in Table 3, respectively for the *noise* dataset and the *random* dataset. In the tables are reported F_1, P and R for the IN class, which is the most important class for this setting, since it represents the users that are re-identified. Most importantly, we report the $\Delta_R(\cdot)$, which is our evaluation metric for testing the change in privacy exposure. This metric reports the difference between the recall of the black-box models w.r.t. the DT, as well as the difference between the recall of the black-box models with respect to the corresponding TREPAN models. In this setting the recall of the IN class is particularly important since it describes how many training records we can reconstruct. A positive value for $\Delta_R(\cdot)$ means that the privacy exposure of the DT or of the TREPAN model is higher w.r.t. the black-box models.

For the MIA attack, we can notice that we have a higher privacy exposure for the DT and the TREPAN models w.r.t. the black-boxes in both of the settings, i.e. *noise* and *random*. The only exception to this trend is $\Delta_R(\text{DT} - \text{RF})$ for BANK, in which the RF has a higher privacy exposure w.r.t. the DT, even if for a small amount. The negative values for this metric may be due to the poorer performance of the DT in this setting which make also the attack less robust. Regarding the privacy attacks against the black-boxes, it is possible to see that overall in the *random* setting the privacy treats are smaller w.r.t. the *noise* one. In particular, the privacy exposure of NN in the *random* case is insignificant (the highest recall in this setting for NN is 0.33 for SYNTH).

The same trends presented for MIA are also present for LABELONLY, especially for the *noise* setting, even if the $\Delta_R(\cdot)$ show lower values w.r.t. the MIA. This result is due to the fact that this attack is better in attacking the black-box models w.r.t. MIA, hence, there is already a higher privacy exposure when attacking the black-boxes. As a consequence, when comparing the difference in privacy exposure between black-boxes and explainers, we find a smaller difference. Regarding the *random* case, LABELONLY obtains a higher privacy exposure w.r.t. MIA, highlighting that this attack is more robust and hence able to succeed even in more difficult settings. However, in the *random* case, LABELONLY shows a decrease in performance w.r.t. the *noise* case. In particular, LABELONLY can attack, with a worrying privacy exposure, both the NN and the RF for the SYNTH dataset. The trend of a higher privacy exposure for the black-boxes in LABELONLY w.r.t. MIA can be seen also for the RF of BANK and for the NN of ADULT, even if in a smaller way.

The performance of ALOA are similar to the LABELONLY for the *noise* case. We can notice similar results also in the case of the $\Delta_R(\cdot)$: the values are lower w.r.t. MIA, but all the attacks show a privacy exposure, both for the explainers and the black-boxes, the firsts higher than the latter. For the *random* setting, instead, ALOA shows a higher privacy exposure w.r.t LABELONLY and MIA. This result was expected due to the procedure of the attack. In fact, ALOA does not require any kind of background knowledge about the original training dataset, not even the statistics of it. Therefore, having privacy exposure with this attack highlights an even more dangerous setting since the attacker can perform it with the only assumption of knowing the shape of the input data, which is public information for on-demand services.

Figure 2 reports the Critical Difference Plot, showing the overall ranking of the three different attacks against global surrogate-based explainers and their black-boxes, both in the *noise* and *random* settings. From this plot, we can observe that there is no statistical difference among the attacks, showing an overall threatening setting for the privacy of the people in the training datasets. Regarding the ranking, ALOA against the TREPAN models is the one which exposes the highest privacy risk, followed by MIA and LABELONLY against TREPAN. However, the three methods against the global explainers have close values in the ranking, with a clear separation between them and the attacks against the black-boxes. In fact, all the attacks against the black-boxes are less powerful w.r.t. the attacks against the explainers, even if the level of privacy exposure remains high. For the attacks against the black-boxes, the ranking is: ALOA, LABELONLY and MIA, but the last one is the lower rank, significantly separated from ALOA and LABELONLY.

5.4 Results attacking the Local Explainers

In this setting we attack the local surrogate explainers. Differently from the global case, the local surrogate is a simple ML model which describes the behaviour of the black-box model close to the record under analysis and not the overall behaviour, as in the case of the global explainers. For this reason, we apply a different procedure, presented in Section 4.1.2. In this setting, the procedure works as follows: firstly, in the *Attack training* procedure, we fit an attack for each surrogate model created ($E = \{e_1, e_2, ..., e_n\}$) together with the attack against the black-box model *b*, obtaining $A_E(\cdot)$, $A_b(\cdot)$. Then, in the *Attack application* procedure, we consider the resulting attack models as part of an ensemble classification method, having $A_E(\cdot)$ as an ensemble of different attacks. The last procedure, *Attack evaluation*, can be instantiated in different ways, depending on the attack considered. In 4.1, we presented the different instantiation of REVEAL in case the attack produces the prediction probabili-

Table 4 Results of the application of three privacy attacks (Mia, Aloa, LabelOnly) with the setting *noise* for training the shadow models, attacking the explainer locally.

-		MIA		LABELONLY		ALOA	
Dataset	**Metric**	RF	NN	RF	NN	RF	NN
Adult	$F1_{\text{In}}$	0.60 (0.00)	0.43 (0.02)	0.77 (0.23)	0.28 (0.10)	0.74 (0.00)	0.45 (0.02)
	P_{In}	0.54 (0.02)	0.30 (0.00)	0.78 (0.78)	0.73 (0.12)	0.79 (0.02)	0.51 (0.06)
	R_{In}	**0.68** (0.02)	**0.70** (0.02)	**0.75** (0.21)	**0.30** (0.10)	**0.72** (0.02)	**0.40** (0.03)
	ΔR_{In}	**0.01**	**0.17**	**-0.06** (0.01)	**-0.37** (0.04)	**-0.08** (0.01)	**-0.13** (0.04)
	Metric	RF	NN	RF	NN	RF	NN
Synth	$F1_{\text{In}}$	0.73 (0.03)	0.70 (0.02)	0.72 (0.02)	0.75 (0.07)	0.73 (0.00)	0.62 (0.00)
	P_{In}	0.71 (0.01)	0.66 (0.00)	0.68 (0.03)	0.65 (0.08)	0.70 (0.01)	0.60 (0.01)
	R_{In}	**0.84** (0.02)	**0.70** (0.00)	**0.78** (0.01)	**0.82** (0.02)	**0.76** (0.02)	**0.65** (0.00)
	ΔR_{In}	**0.02**	**0.18**	**-0.04**	**0.08**	**-0.10** (0.01)	**-0.25** (0.04)
	Metric	RF	NN	RF	NN	RF	NN
Bank	$F1_{\text{In}}$	0.77 (0.01)	0.69 (0.01)	0.58 (0.02)	0.50 (0.00)	0.65 (0.01)	0.43 (0.01)
	P_{In}	0.64 (0.02)	0.68 (0.03)	0.66 (0.05)	0.64 (0.01)	0.58 (0.02)	0.47 (0.00)
	R_{In}	**0.83** (0.00)	**0.69** (0.00)	**0.52** (0.00)	**0.48** (0.05)	**0.71** (0.04)	**0.58** (0.09)
	ΔR_{In}	**0.03**	**0.44**	**-0.24**	**0.29**	**-0.07** (0.01)	**-0.20** (0.04)

Table 5 Results of the application of three privacy attacks (Mia, Aloa, LabelOnly) with the setting *rand* for training the shadow models, attacking the explainer locally.

-		MIA		LABELONLY		ALOA	
Mia	**Metric**	RF	NN	RF	NN	RF	NN
Adult	$F1_{\text{In}}$	0.41 (0.00)	0.46 (0.03)	0.67 (0.23)	0.34 (0.10)	0.69 (0.02)	0.42 (0.05)
	P_{In}	0.30 (0.00)	0.70 (0.01)	0.77 (0.25)	0.73 (0.12)	0.75 (0.02)	0.50 (0.01)
	R_{In}	**0.64** (0.01)	**0.35** (0.04)	**0.60** (0.21)	**0.20** (0.10)	**0.66** (0.01)	**0.38** (0.01)
	ΔR_{In}	**0.28**	**0.03**	**+0.25**	**-0.22**	**0.12**	**-0.02**
	Metric	RF	NN	RF	NN	RF	NN
Synth	$F1_{\text{In}}$	0.64 (0.03)	0.38 (0.00)	0.71 (0.02)	0.63 (0.01)	0.70 (0.01)	0.59 (0.03)
	P_{In}	0.60 (0.01)	0.35 (0.04)	0.65 (0.04)	0.70 (0.01)	0.71 (0.04)	0.60 (0.03)
	R_{In}	**0.71** (0.02)	**0.38** (0.01)	**0.76** (0.02)	**0.60** (0.00)	**0.70** (0.00)	**0.60** (0.01)
	ΔR_{In}	**-0.07**	**+0.05**	**0.00**	**-0.17**	**0.00**	**-0.10**
	Metric	RF	NN	RF	NN	RF	NN
Bank	$F1_{\text{In}}$	0.16 (0.05)	0.46 (0.03)	0.52 (0.03)	0.47 (0.00)	0.68 (0.04)	0.45 (0.04)
	P_{In}	0.27 (0.02)	0.34 (0.09)	0.65 (0.20)	0.60 (0.00)	0.65 (0.00)	0.36 (0.00)
	R_{In}	**0.12** (0.01)	**0.90** (0.01)	**0.50** (0.00)	**0.44** (0.01)	**0.69** (0.02)	**0.60** (0.00)
	ΔR_{In}	**-0.53**	**-0.13**	**-0.23**	**-0.02**	**-0.09**	**-0.14**

ties vector or not. In these experiments, we use the approach which exploits the probability vectors for MIA, while we exploit the other approach for LABELONLY and ALOA.

For the experiments conducted, for each dataset considered we select a set of records to explain from the test set exploiting a K-means clustering procedure, with k being the best value for the dataset under analysis. The choice of k is done by exploiting the elbow method. Due to computational limitations, we explain 3 records for each quantile of each cluster. Regarding the training of the local surrogate models obtained with LORE, the procedure requires synthetically generating a local neighbourhood around the record x under analysis and then fitting a local surrogate DT on the generated neighbours. Therefore, there are different parameters to set. In particular, for this setting, the kind of generation of the neighbourhood and the number of synthetic records to create are important. We conducted a search on these variables, obtaining similar results when considering the *genetic and random* generation, *genetic and probabilities* generation and the *counterfactual first search* generation. The other kinds of generations, such as the *random* one, show lower performance. Regarding the number of synthetic records to create, we use 10000. Similarly to the case of the global explainers, for each local surrogate model we train a MIA attack, with 6 shadow models, with accuracy above 80%. Also in this case, the models created for the attack are all RF. The same setting is applied to LABELONLY and ALOA.

The results of the attacks against the local explainers are reported in Table 4 and in 5, respectively for the *noise* dataset and the *random* dataset. In this setting we observe a lower privacy exposure of the explainers w.r.t. the global setting. This result can be observed by analyzing the values of the $\Delta R(\cdot)$: while in the global case we mostly have positive values, highlighting an increase in privacy exposure when attacking the explainers instead of the black-boxes, in the local case the values are closer to zero, with some negative values, implying that attacking the black-boxes produces a higher privacy exposure than attacking the local explainers. Regarding the *noise* case, MIA produces the highest privacy exposure, with positive $\Delta R(\cdot)$ for all the configurations considered. However, the setting changes in the *random* case, having a lower privacy exposure for MIA and LABELONLY. ALOA, instead, gives similar results both for the *noise* and *random* case, highlighting once again that this attack is the more robust among the three. This result can also be seen in Figure 3, which presented a Critical Difference plot for the Recall of the various attacks performed against the local explainers and their black-boxes. Also, in this case, as in the global case, there is no significant statistical difference among the attacks presented. However, in this plot we can observe that ALOA and LABELONLY against the black-boxes are the highest in the rank, showing a higher privacy exposure w.r.t. ALOA against LORE and MIA against LORE, which is in the fourth position, equally matched and significantly separated from the first two. MIA ranks the lowest among attacks targeting black-box models, while LABELONLY ranks the lowest among attacks on local explanations. Both show a clear distinction from the top quartile of the rankings.

Analysis on the number of records explained For the *local* setting the attack model A_E is an ensemble of multiple attacks, one against each of the local surrogate models created exploiting LORE. To validate REVEAL, we conducted a set of experiments in which we created a local surrogate model for a set of records selected based on a K-means clustering procedure, explaining 3 records for each quantile of each cluster. In practice, our intuition is that the privacy attack will yield better results as the number of records explained increases. This is because explaining more records implies having more local surrogates, which thus better describes the data space under analysis. Consequently, attacking more local models

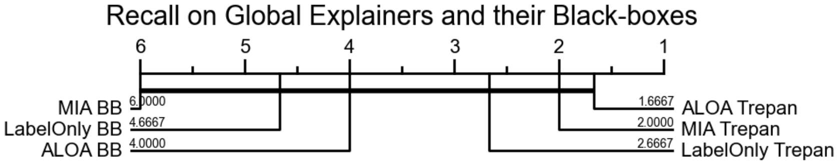

Figure 2: Critical difference plot for the Nemenyi test with $\alpha = 0.05$ for the attacks performed on the *global* explainers (TREPAN) and their associated black-boxes. The reported values result from the ranking procedure and indicate that ALOA, MIA, and LABELONLY against TREPAN show minimal differences among themselves. In contrast, the attacks targeting the black-boxes rank lower and are clearly separated from the top three.

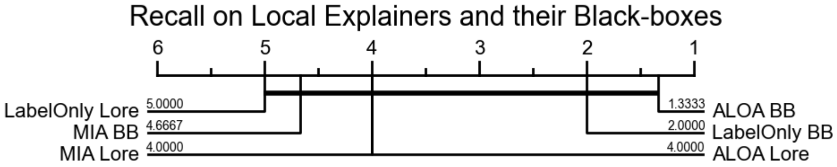

Figure 3: Critical difference plot for Nemenyi test with $\alpha = 0.05$ for the attacks performed on the *local* explainers and their black-boxes. The values reported results from the ranking procedure. ALOA and MIA against the black-boxes are the highest in the rank, posing a higher privacy threat. The ranking values show a clear separation between them and the other attacks.

that better describe the space under analysis should also improve the ensemble method of attacks. To validate this insight, we increase the number of records explained for each cluster. In particular, we consider ALOA with the *noise* dataset for SYNTH, which is the setting that shows a higher privacy exposure, and increase the number of elements for each of the datasets considered, ranging from 40 records up to 120 records. The results are reported in Figure 4. From the plot, it is possible to observe that with a small number of records, the performance of the attack is low, highlighting that with few local explanations, the risk of privacy is low. This result aligns with our expectations, as limited local surrogate availability cannot represent all facets of the data space under analysis. However, the increase in the number of records explained also leads to an increase in privacy exposure, reaching a plateau starting from 80 records explained for all the datasets, i.e., starting from 80 records, the increase in the number of records do not show an increase in privacy exposure. This behavior in the privacy risk analysis is a finding already reported in literature [30] in the setting of assessing the privacy of the data.

6 Conclusion

In this paper we propose REVEAL, a framework for assessing the privacy exposure of the black-box models and their surrogate-based explainers, being them local or global. The method proposed is generic and can be exploited for every kind of black-box model, every surrogate-based explainer and with different kinds of privacy attacks. The analysis conducted shows that attacking the privacy of the explainers, being local or global, gives rise to privacy exposures. Depending on the privacy attack considered, we have different levels of privacy risks, rising to particularly concerning situations with ALOA, a privacy

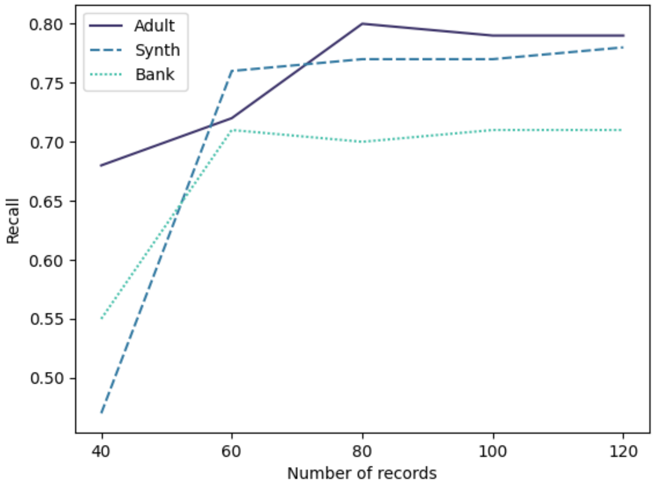

Figure 4: Results of ALOA increasing the number of records explained for the *local* setting. Starting from 80 records explained all the datasets reach a plateau in the privacy exposure.

attack that has very little assumption of knowledge on the part of the attacker and yet still manages to achieve good results in terms of privacy breaches. However, the global explainers show higher privacy exposure with respect to their black-boxes. At the same time, this is not the case for the local explainers, which show the same or lower level of privacy exposure as their corresponding black-boxes.

In this paper we focused our analysis on tabular data due to the limited availability of surrogate-based explainers for more complex data types. However, as a future work, we plan to extend our study to more complex data, such as time-series, for which some explainable AI methods exploit surrogate models (e.g., LASTS [41]). Applying our framework in this context will also require to design and develop privacy attacks specifically tailored to machine learning models for such type of data, which inherently models temporal dependency between a series of observations.

The results obtained evaluating REVEAL expose a concerning scenario in which user privacy is at significant risk, particularly when global explainers are employed. Our findings highlight the delicate balance between explainability and privacy that must be carefully managed in the development of Artificial Intelligence systems. In future work, we aim to explore the generation of explanations that protect user privacy. Recently, there has been growing interest in privacy-protected explanations, as discussed in [4]. For this reason, we plan to assess the privacy implications of differentially private explanations, such as SHAP values, by analyzing both cases where privacy protection methodologies are applied to the data or to the training of the machine learning models to explain.

Acknowledgments. This work is partially founded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - ”FAIR - Future Artificial Intelligence Research” - Spoke 1 ”Human-centered AI”, funded by the European Commission under the NextGeneration EU programme; PNRR -“SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics” - Funded by the European Union; the European Union’s Horizon Europe research and innovation programme for the project FINDHR (g.a. No. 101070212); ERC-2018-ADG G.A. 834756 XAI: Science and technology for the eXplanation of AI decision making (`https://xai-project.eu/index.html`); by the European Union under Grant Agreement no. 101120763 - TANGO. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] M. Al-Rubaie and J. M. Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security Privacy*, 17, 2019.

[2] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knoweledge Discovery*, 37:1719–1778, 2023.

[3] O. Boz. Extracting decision trees from trained neural networks. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[4] A. Bozorgpanah, V. Torra, and L. Aliahmadipour. Privacy and explainability: The effects of data protection on shapley values. In *Technologies 2022*, volume 10, page 125, 2022.

[5] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models, 2021.

[6] China. The china regulation: Internet information service algorithmic recommendation management provisions. Link.

[7] China. The china regulation: New generation artificial intelligence development plan. Link.

[8] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot. Label-only membership inference attacks. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR, 18–24 Jul 2021.

[9] L. Corbucci, A. Monreale, and R. Pellungrini. Enhancing privacy and utility in federated learning: A hybrid p2p and server-based approach with differential privacy protection. In *Proceedings of the 21st International Conference on Security and Cryptography - Volume 1: SECRYPT*, pages 592–602. INSTICC, SciTePress, 2024.

[10] M. W. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In *JMLR*, pages 37–45. Elsevier, 1994.

[11] C. Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Vol. Part II*, ICALP'06, 2006.

[12] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, 2015.

[13] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, 2014.

[14] A. A. Freitas. Comprehensible classification models: a position paper. *SIGKDD Explor.*, 15(1):1–10, 2013.

[15] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, 2018.

[16] F. O. Gomes, R. Pellungrini, A. Monreale, C. Renso, and J. E. Martina. Trajectguard: A comprehensive privacy-risk framework for multiple-aspects trajectories. *IEEE Access*, 12:136354–136378, 2024.

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi, and F. Giannotti. Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Mining and Knowledge Discovery*, 2022.

[18] D. S. W. Jakob Mökander, Prathm Juneja and L. Floridi. The us algorithmic accountability act of 2022 vs. the eu artificial intelligence act: what can they learn from each other? In *Minds & Machines*, volume 32, pages 751–758, 2022.

[19] Japan. Japan: the development of artificial intelligence. Link.

[20] B. Kim, R. Khanna, and O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. NIPS'16, 2016.

[21] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774. 2017.

[22] G. Mariani, A. Monreale, and F. Naretto. Privacy risk assessment of individual psychometric profiles. In *Soares, C., Torgo, L. (eds) Discovery Science*, volume 12986. Springer, Cham., 2021.

[23] A. Monreale, F. Naretto, and S. Rizzo. Agnostic label-only membership inference attack. In *17th International Conference on Network and System Security*. Springer, 2023.

[24] F. Naretto, A. Monreale, and F. Giannotti. Evaluating the privacy exposure of interpretable global explainers. In *4th IEEE International Conference on Cognitive Machine Intelligence, CogMI 2022, Atlanta, GA, USA, December 14-17, 2022*, pages 13–19. IEEE, 2022.

[25] F. Naretto, R. Pellungrini, D. Fadda, and S. Rinzivillo. Exphlot: Explainable privacy assessment for human location trajectories. In *Discovery Science*, volume 14276 of *Lecture Notes in Computer Science*. Springer, Cham., 2023.

[26] F. Naretto, R. Pellungrini, A. Monreale, F. M. Nardini, and M. Musolesi. Predicting and explaining privacy risk exposure in mobility data. In *Discovery Science - 23rd International Conference, DS 2020, Thessaloniki, Greece, October 19-21, 2020, Proceedings*, Lecture Notes in Computer Science. Springer, 2020.

[27] F. Naretto, R. Pellungrini, F. M. Nardini, and F. Giannotti. Prediction and explanation of privacy risk on mobility data with neural networks. In *ECML PKDD 2020 Workshops*. Springer International Publishing, 2020.

[28] 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. *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.

[29] S. C. on Artificial Intelligence. The national artificial intelligence research and development strategic plan: 2019 update. In *Executive Office of the President of the United States*. Curran Associates, Inc., 2019.

[30] F. Pratesi et al. Prudence: a system for assessing privacy risk vs utility in data sharing ecosystems. *Transactions on Data Privacy*, 11(2):139–167, 2018.

[31] ProPublica. The compas recidivism case. Link.

[32] P. Quan, S. Chakraborty, J. V. Jeyakumar, and M. Srivastava. On the amplification of security and privacy risks by post-hoc explanations in machine learning models, 2022.

[33] M. T. Ribeiro et al. "Why should I trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144, 2016.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535, 2018.

[35] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *PODS*, page 188. ACM, 1998.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 2020.

[37] R. Shokri, M. Strobel, and Y. Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[38] 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)*, 2017.

[39] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[40] L. Song, R. Shokri, and P. Mittal. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*, 2019.

[41] F. Spinnato, R. Guidotti, A. Monreale, M. Nanni, D. Pedreschi, and F. Giannotti. Understanding any time series classifier with a subsequence-based explainer. *ACM Trans. Knowl. Discov. Data*, 18(2), Nov. 2023.

[42] U. State. The algorithmic accountability act of united states. Link.

[43] U. State. The blueprint for an ai bill of rights (united states). Link.

[44] V. Torra. *Data Privacy: Foundations, New Developments and the Big Data Challenge*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[45] UK. The uk policy paper: Ai sector deal. Link.

[46] Unesco. The unesco's ethics of artificial intelligence. Link.

[47] E. Union. The artificial intelligence act. Link.

[48] E. Union. The general data protection regulation. Link.