Inferring Sensitive Attributes from Model Explanations

Vasisht Duddu, Antoine Boutet

Inferring Sensitive Attributes from Model Explanations

Vasisht Duddu
University of Waterloo
Waterloo, Canada
vasisht.duddu@uwaterloo.ca

Antoine Boutet
Univ Lyon, INSA Lyon, Inria, CITI
Lyon, France
antoine.boutet@insa-lyon.fr

ABSTRACT

Model explanations provide transparency into a trained machine learning model's blackbox behavior to a model builder. They indicate the influence of different input attributes to its corresponding model prediction. The dependency of explanations on input raises privacy concerns for sensitive user data. However, current literature has limited discussion on privacy risks of model explanations.

We focus on the specific privacy risk of *attribute inference attack* wherein an adversary infers sensitive attributes of an input (e.g., Race and Sex) given its model explanations. We design the first attribute inference attack against model explanations in two threat models where model builder either (a) includes the sensitive attributes in training data and input or (b) censors the sensitive attributes by not including them in the training data and input.

We evaluate our proposed attack on four benchmark datasets and four state-of-the-art algorithms. We show that an adversary can successfully infer the value of sensitive attributes from explanations in both the threat models accurately. Moreover, the attack is successful even by exploiting only the explanations corresponding to sensitive attributes. These suggest that our attack is effective against explanations and poses a practical threat to data privacy.

On combining the model predictions (an attack surface exploited by prior attacks) with explanations, we note that the attack success does not improve. Additionally, the attack success on exploiting model explanations is better compared to exploiting only model predictions. These suggest that model explanations are a strong attack surface to exploit for an adversary.

KEYWORDS

Attribute Privacy, Inference Attacks, Explainable Machine Learning.

1 INTRODUCTION

Machine Learning (ML) models are used for high-stakes decision making for several real-world applications. For instance, these models assist decision makers such as doctors and judges in healthcare and criminal justice [27]. However, the model's high complexity makes it difficult for human interpretation into the decision making process. This creates the need for *transparency* into the model behaviour. Model explanations release additional information to explain the behaviour of complex ML models. Specifically, attribute based model explanations explain the model's prediction on an input by releasing the influence of different input attributes responsible for the prediction [2, 22, 30, 34, 37].

Some of the input attributes can be sensitive (e.g., Race and Sex). This raises the data privacy concerns when an adversary ($\mathcal{A}dv$) can leverage model explanations as an attack surface. For instance, Shokri et al. [29] show that explanations can be exploited for membership inference (i.e., inferring whether input record was part of training data) and data reconstruction. Additionally, releasing model explanations could leak the values of sensitive attributes which is a privacy risk, not considered in literature. For instance, consider the setting where an ML model is trained to predict the likelihood that a criminal will re-offend as an aid to judges in a court. In addition to output predictions, the model reveals explanations on why it made the prediction on that input. Attribute inference attacks could reveal Race and Sex from model explanations which individual prefers to keep their private to avoid biased decisions.

However, this quantification of privacy risk of model explanations to *attribute inference attacks* is lacking in current literature. An analysis of this trade-off between privacy and transparency is necessary so that a model builder ($\mathcal{M}$) can make appropriate choices to train ML models for high-stakes applications. In this work, we ask the following research question: *can an $\mathcal{A}dv$ exploit model explanations to infer sensitive attributes of individual data records?* We design the *first attribute inference attack* to infer sensitive attributes from model explanations in two threat models:

TM1 Sensitive attributes are included in the training dataset and the input (following prior work [8, 9]) and $\mathcal{A}dv$ only sees the output predictions but not their inputs. $\mathcal{A}dv$ has no control over passing the inputs but has to infer sensitive attributes from only the observed predictions.

TM2 Sensitive attributes are not included in training data or input (censored by $\mathcal{M}$ for privacy). This corresponds to real-world application such as ML as a Service (MLaaS).

In this work, we claim the following main contributions.

(1) We design the **first** attribute inference attack, to infer sensitive attributes, e.g., Race and Sex, of the data records from corresponding model explanations. $\mathcal{A}dv$ trains an ML attack model to map model explanations to sensitive attributes. We additionally calibrate the threshold over the attack model's predictions to increase $\mathcal{A}dv$'s power (Section 4).

(2) In **TM1**, we show that our attack successfully infers the sensitive attributes from model explanations (Section 6). On evaluating across four benchmark datasets and four model explanations, we note:
   - a high F1-score of 0.92 ± 0.07 (Race) and 0.88 ± 0.11 (Sex) using entire model explanations corresponding to both sensitive and non-sensitive attributes (Section 6.1).
   - a high F1-score of 0.90 ± 0.10 (Race) and 0.83 ± 0.09 (Sex) using model explanations corresponding to only sensitive attribute (Section 6.2).

(3) In **TM2**, despite censoring the sensitive attributes, we show that our attack can successfully infer them using model explanations of other non-sensitive attributes (Section 7). On evaluating across four benchmark datasets and four model explanations, we note:
   - a high F1-score of 0.83 ± 0.12 (Race) and 0.77 ± 0.09 (Sex) (Section 7.1).

- that on combining model explanations with model predictions, attack success does not improve. Hence, model explanations are a strong attack surface for $\mathcal{A}dv$ to exploit (Section 7.2).

(4) In both **TM1** and **TM2**, exploiting model explanations has a higher success than prior state-of-the-art attribute inference attacks which exploit model predictions. This indicates that releasing model explanations increases the attack surface enabling $\mathcal{A}dv$ to mount strong attribute inference attacks (Section 8).

2 BACKGROUND

Consider a training dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{S}, \mathcal{Y}\}$ where $\mathcal{X}$ is the space of non-sensitive input attributes, $\mathcal{S}$ is the space of sensitive input attributes, $\mathcal{Y}$ is the space of classification labels. We denote a data record as (x, y, s) with non-sensitive attributes x and sensitive attribute s where $(x, s) \in \mathcal{X} \times \mathcal{S}$ and classification label $y \in \mathcal{Y}$. ML models learn a function $f_\theta : (x \cup s) \rightarrow y$ which maps the input with sensitive and non-sensitive attributes to y. Alternatively, the models can be trained without s in the training dataset given by $f_\theta : x \rightarrow y$. The models are parameterized by θ which are iteratively updated to minimize the loss on correctly predicting x or $x \cup s$ as y. The model training, hyperparameters selection and deployment to application is done by $\mathcal{M}$.

Given these formal notations, we describe the state-of-the-art algorithms for model explanations considered in this work (Section 2.1) and prior work on attribute inference attacks (Section 2.2).

2.1 Model Explanations

Model explanations describe a model's behaviour to $\mathcal{M}$ on specific inputs. Specifically, attribute based model explanations estimate the influence of input attributes on the model's output prediction. In other words, these explanations assign a score to each attribute in the input point of interest (PoI) which resulted in a particular model prediction. Formally, for a given PoI $\vec{x} = (x_1, \cdots x_n)$, the model explanations $\phi(\vec{x})$ outputs a vector indicating the importance of different attributes influential in the model's prediction of $\vec{x}$. Here, $\phi(\vec{x})$'s attribution of the prediction at input PoI $\vec{x}$ relative to a baseline input $\vec{x'}$ is a vector $\phi_{\vec{x'}}(\vec{x}) = (\phi_1, \cdots, \phi_n)$.

We consider two types of attribute based explanation algorithms: (a) backpropagation-based explanations (IntegratedGradients and DeepLift) and (b) perturbation-based explanations (GradientSHAP and SmoothGrad).

Gradient-based Explanations compute gradients using backpropagation to estimate the influence of attributes to predictions.

- **IntegratedGradients** [37] computes the integration of gradients with respect to inputs by considering a straight line path from the baseline $\vec{x'}$ to the PoI $\vec{x}$. This integration across the i^{th} dimension can be computed as: $\phi_{\text{IntegratedGradients}_i}(\vec{x}) = (\vec{x} - \vec{x'}) \times \int_{\alpha=0}^{1} \frac{\partial f_\theta(\vec{x'} + \alpha(\vec{x} - \vec{x'}))}{\partial x_i} d\alpha$. Here, $\frac{\partial f_\theta(x)}{\partial x_i}$ indicates the gradient computed using the model f_θ over the input x across the i^{th} dimension.
- **DeepLift** [2, 30] estimates the contribution of specific neurons using the difference in output with respect to a baseline output. It assignes scores as $\phi_{\text{DeepLift}}(\vec{x}) = m_{\Delta\vec{x}\Delta t} \frac{C_{\Delta\vec{x}\Delta t}}{\Delta\vec{x}}$ where x is a given input neuron, Δx is the difference from baseline, t is target neuron and its output difference from baseline is given as Δt. The multiplier captures the contribution of Δx to Δt and is similar to partial derivative but over finite differences instead of infinitesimal differences.

Perturbation-based Explanations add noise to data records or remove some attributes to see the impact on the model utility.

- **GradientSHAP** [22] computes the Shapley values and adds Gaussian noise to each input PoI by sampling multiple times and selects a random input x along the path between baseline and input. The gradient of outputs with respect to those selected random points are then computed. In other words, the final attributes are computed as the expected value over the product of the gradient and the difference in input PoI to the baseline: $\frac{\partial f_\theta(x)}{\partial x} \times (\vec{x} - \vec{x'})$.
- **SmoothGrad** [34] samples random inputs in a neighborhood of PoI $\vec{x}$ by adding Gaussian noise to the PoI. Then it averages the resulting sensitivity maps (i.e., derivative of model predictive with respect to input) corresponding to the n noisy neighbour records $\phi_{\text{SmoothGrad}}(\vec{x}) = \frac{1}{n} \sum_1^n \frac{\partial f_\theta(\vec{x} + \mathcal{N}(0, \sigma^2))}{\vec{x}}$.

A natural choice for baseline $\vec{x'}$ to compute model explanations is where the prediction is unbiased [37]. In all the cases, we use the mean vector over the inputs as our baseline. Additionally, each model explanation algorithm also outputs a convergence delta, δ where the lower the absolute value of the convergence delta the better is the approximation (i.e., low error). We append δ with $\phi()$ to obtain the final attack vector. We abuse the notation to refer the appended vector as $\phi()$.

2.2 Attribute Inference Attacks

Attribute inference attacks aim to infer s (e.g., $s = 1$ for males and $s = 0$ for females) for an individual data record. $\mathcal{A}dv$ exploits observable information (i.e., model predictions or explanations in our case) to infer unobservable information (i.e., s). This attack is different from property inference attacks proposed in literature which aim to infer global properties of dataset (e.g., inferring the ratio of males to female attributes on which the model was trained on) [10, 24, 42].

Several prior work have proposed attribute inference attacks against ML models using the model's output predictions [7–9, 23, 35, 40]. Fredrikson et al. [8, 9] propose an attribute inference attack where $\mathcal{A}dv$ infers s using the knowledge of both x and $f_\theta(x \cup s)$. However, this assumption of $\mathcal{A}dv$'s knowledge is strong. Mahajan et al. [7] and Song et al. [35] proposed an attack where an ML attack model was trained to infer s using only model prediction. This attack exploits the distinguishability in predictions conditioned on different values of s. However, the attack model performs poorly for imbalanced dataset since the default threshold of 0.5 for estimating the value of s is incorrect for skewed prediction distribution. To address this, Aalmoes et al. [1] proposed an attack which accounts for this skewness of attack model's predictions. They select a threshold over attack model's predictions which maximizes attack F1-Score on an auxiliary dataset known to $\mathcal{A}dv$. That threshold is used over attack model's predictions to infer s for target data records.

3 PROBLEM STATEMENT

Our goal is to evaluate the privacy risks of model explanations to attribute inference attacks and hence study the trade-offs between privacy and transparency. We consider the following setting: target ML model f_{target} is trained and deployed on the Cloud by $\mathcal{M}$ within MLaaS paradigm. Given a POI $\vec{x}$, we assume that f_{target} can output both the model prediction ($f_{target}(\vec{x})$) and corresponding explanations on that input ($\phi(\vec{x})$). $\phi()$ are required to be released by AI regulations to ensure trustworthy computation [14, 15, 17, 21, 38].

Given that model explanations measure the influence of individual attributes in the input to the model's prediction, it is natural to ask, *given access to $\phi()$, can $\mathcal{A}dv$ infer s?* This study is currently lacking in literature. We describe three main requirements for the design of an effective attack:

AR1 Attack should operate in a **blackbox threat model**, where $\mathcal{A}dv$ sends an input and obtains an output via an API from a MLaaS service provider. $\mathcal{A}dv$ does not have access to f_{target}'s internal parameters or architecture.

AR2 Attack should be **practical**, i.e., uses model observables ($\phi()$ or $f_{target}()$) to infer unobservables (s).

AR3 Attack should **account for class imbalance in** s. In all practical applications, s is imbalanced which skews the predictions of f_{adv} lowering the $\mathcal{A}dv$'s attack success to correctly infer s.

AR4 Attack should be **applicable to model explanations**, i.e., exploits $\phi()$ to infer the values of s.

Prior attacks exploit the distinguishability in $f_{target}()$ given different values of s [1, 7–9, 35, 40]. Fredrikson et al. [8, 9] and Yeom et al. [40] attacks have strong assumptions about $\mathcal{A}dv$ knowledge, such as knowledge of x in addition to $f_{target}()$ (violating requirement **AR2**). Alternatively, Song et al. [35] and Mahajan et al. [7] do not account for the class imbalance in s (violating requirement **AR3**). Aalmoes et al. [1] use threshold calibration to improve the attack success but they exploit $f_{target}()$ and not $\phi()$ (violating **AR4**).

3.1 Threat Model and Attack Methodology

We discuss two threat models **TM1** and **TM2** along with the assumptions about $\mathcal{A}dv$'s knowledge and attack methodology.

- **TM1 (w/ s in $\mathcal{D}$):** We assume s is included in both $\mathcal{D}$ and input (i.e., $x \cup s$). Hence, $\phi(x \cup s)$ are released as part of the API and $\mathcal{A}dv$ can obtain $\phi(s)$ along with $\phi(x)$. This is when the $\mathcal{A}dv$ is monitoring the outputs from the model. Here, $\mathcal{A}dv$ cannot choose inputs to send to f_{target} (as it already includes s)[1] but can only observe $f_{target}(x \cup s)$ and $\phi(x \cup s)$ for some arbitrary inputs. Given only $\phi(x \cup s)$, $\mathcal{A}dv$ aims to infer s using an attack ML model $f_{adv} : \phi(x \cup s) \rightarrow s$. Here, $\mathcal{A}dv$ can also attack different model explanations: $f_{adv} : \phi(x) \rightarrow s$ and $f_{adv} : \phi(s) \rightarrow s$[2].
- **TM2 (w/o s in $\mathcal{D}$):** We assume s is *not* included in $\mathcal{D}$ and input (i.e., x). $\mathcal{A}dv$ has blackbox access to f_{target} and pass an input x and obtain access to both $\phi(x)$ and $f_{target}(x)$. Unlike **TM1**, $\mathcal{A}dv$ can choose the input to pass to the model. This is the worst case for $\mathcal{A}dv$ where s is censored by $\mathcal{M}$ for privacy making this threat model more practical. Given $\phi(x)$, $\mathcal{A}dv$ aims to infer s using an attack ML model $f_{adv} : \phi(x) \rightarrow s$.

[1]Despite this, this is seen in several prior attribute inference attacks [7–9, 35, 40].
[2]$\phi(s)$ and $\phi(x)$ indicates model explanations corresponding to s and x respectively.

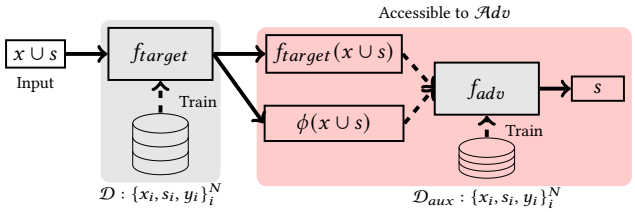


Figure 1: TM1 threat model: train f_{target} on training data with s included. $\mathcal{A}dv$ only has access to predictions $f_{target}(x \cup s)$ and explanations $\phi_{target}(x \cup s)$ but cannot pass inputs. Attack requires training f_{adv} on $\mathcal{D}_{aux}$ to infer s given $\phi_{target}(x \cup s)$.

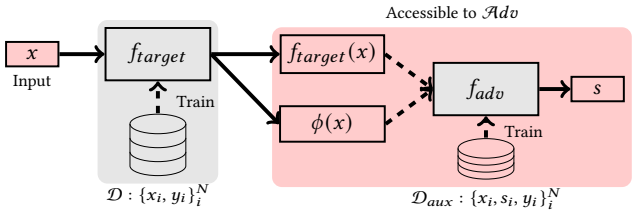


Figure 2: TM2 threat model: train f_{target} on data without s. $\mathcal{A}dv$ has access to predictions $f_{target}(x)$ and explanations $\phi_{target}(x)$ and choose the inputs to pass (w/o s). Attack trains f_{adv} on $\mathcal{D}_{aux}$ to infer s given $\phi_{target}(x)$.

Figure 1 and 2 where red indicates components accessible to $\mathcal{A}dv$. In both **TM1** and **TM2**, we assume that $\mathcal{A}dv$ has additional auxiliary dataset $\mathcal{D}_{aux}$ which is drawn from the same distribution as $\mathcal{D}$ and includes data records (x, s, y) containing non-sensitive and sensitive attributes with corresponding label. This assumptions is inline with all the prior attribute inference attacks proposed in literature [7–9, 35, 40]. $\mathcal{D}_{aux}$ is used to train f_{adv}: $\mathcal{A}dv$ passes data records (x, s, y) (**TM1**) or (x, y) (**TM2**) to f_{target} and uses the generated model explanations to train f_{adv} by mapping them to s (known to $\mathcal{A}dv$ for $\mathcal{D}_{aux}$). This access to f_{target} can be alleviated by training a "shadow model" on $\mathcal{D}_{aux}$ to mimic f_{target}; and use the model explanations from "shadow model" to train f_{adv}. We only use predictions from f_{target} to train f_{adv}. Once f_{adv} is trained, the attack is evaluated on target dataset (distinct from $\mathcal{D}_{aux}$).

Moreover, any attack designed within **TM1** and **TM2** are blackbox (satisfy **AR1**) and practical (satisfy **AR2**). In both threat models, $\phi()$ is accessible to adversary to exploit and hence satisfies **AR4**. Given this, we now have to design an attack which satisfies requirement **AR3** to account for class imbalance in s and improve attack success.

4 OUR PROPOSED ATTACK

Prior attribute inference attacks are directly applicable as they do not satisfy requirements **AR1**-**AR4**. We design attribute inference attacks to adapt to $\phi()$ to infer s while calibrating the threshold over f_{adv}'s predictions to improve attack success. Instead of using the

default threshold of 0.5, as in prior attacks over model predictions [7, 35], we calibrate the threshold over $f_{adv}(\phi())$ to maximize F1-Score. We compute an optimal threshold τ^* over the probability $P(s|\phi(x))$, which is the output of $f_{adv}(\phi(x))$, to infer s. In practice, we use the precision-recall curve which computes precision and recall values for multiple thresholds over f_{adv}'s predictions. Then, τ^* is chosen based on maximum F1-Score and this in-turn improves the precision and recall values. This is effective when there is a moderate to large class imbalance (satisfies **AR3**).

Calibrating the Threshold. First, as a sanity check, we ensure the precision-recall curves are above random guess baseline. A random guess for precision-recall curve is the horizontal line with the precision value computed over the positive class examples in the dataset. Figure 3 shows the precision-recall curves for f_{adv} on $\mathcal{D}_{aux}$ which is beyond random guess in all cases. This indicates the possibility of finding τ^* to improve $\mathcal{A}dv$'s F1-Score.

Table 1: τ^* is different from default threshold of 0.5. "IG" is IntegratedGradients, "DL" is DeepLift, "GS" is GradientSHAP and "SG" is SmoothGrad.

	IG				DL			
	w/ S		w/o S		w/ S		w/o S	
Dataset	Race	Sex	Race	Sex	Race	Sex	Race	Sex
CENSUS	0.64	0.47	0.42	0.54	0.96	0.51	0.82	0.37
COMPAS	0.94	0.89	0.38	0.59	0.97	0.84	0.38	0.52
LAW	0.93	0.56	0.93	0.56	0.93	0.74	0.79	0.56
CREDIT	0.55	0.42	0.54	0.48	0.61	0.55	0.46	0.40
	GS				SG			
	w/ S		w/o S		w/ S		w/o S	
Dataset	Race	Sex	Race	Sex	Race	Sex	Race	Sex
CENSUS	0.77	0.26	0.55	0.47	0.68	0.49	0.51	0.50
COMPAS	0.61	0.58	0.46	0.54	0.81	0.72	0.33	0.55
LAW	0.68	0.61	0.82	0.56	0.97	0.96	0.93	0.57
CREDIT	0.48	0.48	0.56	0.52	0.60	0.43	0.51	0.44

Table 1 further shows that the resultant τ^* is indeed different from 0.5 default threshold, indicative of improving attack success. It is important to note that τ^* is computed on $\mathcal{D}_{aux}$ which might be different from optimal threshold on target dataset which is being attacked. However, this cannot be known before-hand by $\mathcal{A}dv$. This is the best $\mathcal{A}dv$ can do before performing the attack in real-world with imbalanced datasets hoping that τ^* improves attack success.

Why is this a Privacy Risk? Our threat models are similar to prior attacks [7–9, 35, 40]. One can argue that the attack is not actually exploiting the model explanations but using the existing correlations between sensitive and non-sensitive attributes (which $\mathcal{A}dv$ could deduce from $\mathcal{D}_{aux}$). In this case, there is no privacy violation. However, as seen in Table 2, Pearson's correlation between s and other attributes is low. Hence, $\mathcal{A}dv$ exploits non-trivial information, i.e., information memorized by f_{target} about s which is present in $\phi()$ (similar to the case of inferring s from $f_{target}()$ [35]).

Table 2: Low Pearson Correlation of s with y, x, $\phi(s)$ and $\phi(x)$ indicates that model is memorizing unintended private data.

Dataset	y		x	
	Race	Sex	Race	Sex
CENSUS	0.02	0.01	0.00 ± 0.02	0.00 ± 0.02
COMPAS	-0.06	0.02	-0.01 ± 0.03	-0.02 ± 0.05
LAW	0.02	0.02	-0.01 ± 0.02	0.00 ± 0.01
CREDIT	0.01	-0.01	0.01 ± 0.01	0.00 ± 0.02
IntegratedGradients				
Dataset	$\phi(s)$		$\phi(x)$	
	Race	Sex	Race	Sex
CENSUS	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02
COMPAS	-0.01 ± 0.02	-0.01 ± 0.06	-0.01 ± 0.02	0.00 ± 0.07
LAW	0.02 ± 0.00	0.02 ± 0.02	0.02 ± 0.00	0.01 ± 0.02
CREDIT	0.01 ± 0.02	0.00 ± 0.03	0.01 ± 0.02	0.00 ± 0.02
DeepLift				
Dataset	$\phi(s)$		$\phi(x)$	
	Race	Sex	Race	Sex
CENSUS	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02
COMPAS	0.00 ± 0.03	-0.02 ± 0.06	-0.01 ± 0.02	0.01 ± 0.06
LAW	-0.01 ± 0.03	-0.00 ± 0.01	-0.02 ± 0.03	0.00 ± 0.01
CREDIT	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.02	0.00 ± 0.02
GradientSHAP				
Dataset	$\phi(s)$		$\phi(x)$	
	Race	Sex	Race	Sex
CENSUS	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02
COMPAS	-0.01 ± 0.04	-0.01 ± 0.03	-0.03 ± 0.02	-0.01 ± 0.03
LAW	-0.01 ± 0.02	0.00 ± 0.00	-0.02 ± 0.00	0.00 ± 0.00
CREDIT	0.00 ± 0.01	0.01 ± 0.02	0.00 ± 0.01	0.01 ± 0.02
SmoothGrad				
Dataset	$\phi(s)$		$\phi(x)$	
	Race	Sex	Race	Sex
CENSUS	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02	0.00 ± 0.02
COMPAS	0.00 ± 0.02	-0.01 ± 0.07	-0.01 ± 0.02	0.00 ± 0.08
LAW	-0.04 ± 0.04	-0.01 ± 0.03	-0.03 ± 0.06	0.01 ± 0.03
CREDIT	0.01 ± 0.02	0.00 ± 0.02	0.01 ± 0.02	0.00 ± 0.02

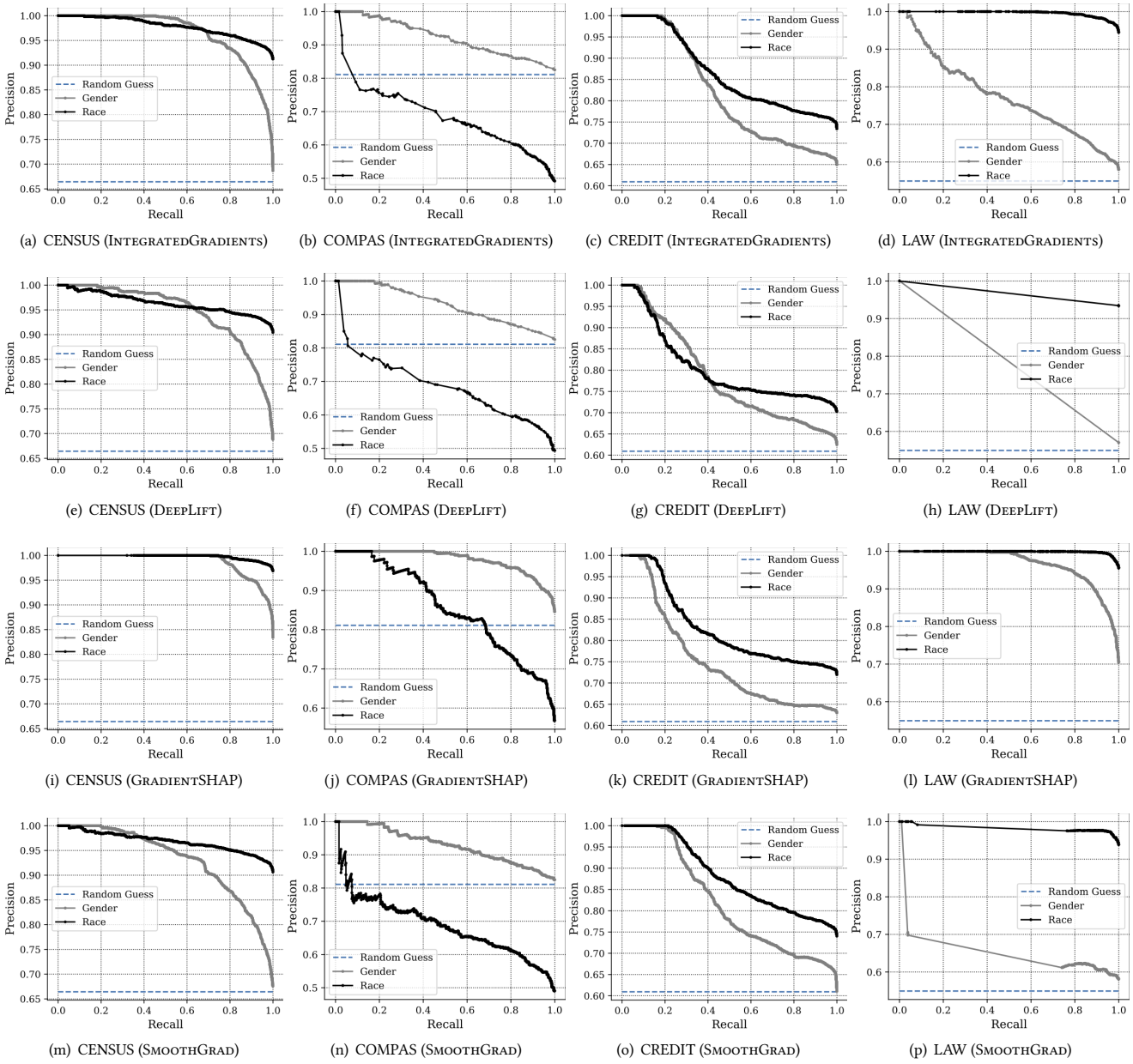
5 EXPERIMENTAL SETUP

We describe the tabular benchmark datasets used in our evaluation (Section 5.1), f_{target} and f_{adv} architectures (Section 5.2), and metrics for evaluating attack success (Section 5.3).

5.1 Datasets

We consider four tabular datasets to demonstrate different high-stakes decision making applications. All the datasets have a binary classification task and publicly available.

- **Adult Income (CENSUS)** comprises of 48,842 data records with 14 attributes about individuals from 1994 US Census data. The attributes include marital status, education, occupation, job hours per week among others. The binary classification is whether an individual makes an income of 50k per annum.



(a) CENSUS (IntegratedGradients) (b) COMPAS (IntegratedGradients) (c) CREDIT (IntegratedGradients) (d) LAW (IntegratedGradients)

(e) CENSUS (DeepLift) (f) COMPAS (DeepLift) (g) CREDIT (DeepLift) (h) LAW (DeepLift)

(i) CENSUS (GradientSHAP) (j) COMPAS (GradientSHAP) (k) CREDIT (GradientSHAP) (l) LAW (GradientSHAP)

(m) CENSUS (SmoothGrad) (n) COMPAS (SmoothGrad) (o) CREDIT (SmoothGrad) (p) LAW (SmoothGrad)

Figure 3: Precision-recall curves for finding optimal threshold to improving $\mathcal{A}dv$'s success. The precision recall are above random guess which can allow $\mathcal{A}dv$ to compute an optimal threshold to improve attack success.

- **Recidivism (COMPAS)** is used for commercial algorithms by judges and parole officers for estimating the likelihood of a criminal reoffending. It contains 10,000 criminal defendants in Florida. The binary classification is if a criminal will reoffend.
- **Law School Dataset (LAW)** is based on survey conducted by Law School Admission Council across 163 law schools in the United States. It contains information on 21,790 law students such as their entrance exam scores (LSAT), their grade-point average (GPA) collected prior to law school, and their first year average grade. The classification is to predict if an applicant will have a high first year average grade.
- **UCI Credit Card (CREDIT)** is an anonymized dataset from the UCI Machine Learning dataset repository and contains information about different credit card applicants. The dataset contains 30,000 records with 24 attributes for each record. The binary classification is if the application was approved.

In all the four datasets, the sensitive attributes are Race and Sex. We use this sensitive attributes for demonstration. The attack

however can extend to other sensitive attributes with discrete values. We use 70% of $\mathcal{D}$ for f_{target} and the remaining 30% as testing dataset. $\mathcal{D}_{aux}$ is 50% of the testing dataset while the other half is used as unseen dataset for evaluating the attack success.

5.2 Architecture

We now describe the model architectures and training hyperparameters for f_{target}, trained on the main classification task, and f_{adv} used by $\mathcal{A}dv$ to map the explanations to s. We use pytorch and captum library for model explanations and our code is made publicly available: https://github.com/vasishtduddu/AttInfExplanations.git.

- **Target Models.** We consider a fully connected neural network with four hidden layers of sizes [1024, 512, 256, 128] for all the datasets. Note that all the datasets have a binary classification tasks and the target models are binary classifiers. The target models are trained for 30 epochs with Adam optimizer and learning rate of 1e-3 and no regularization.
- **Attack Models.** We use a neural network model for all datasets other than LAW, where we use a random forest classifier with a maximum depth of 150. We consider a fully connected neural network with three hidden layers of sizes [64, 128, 32]. The model is trained using Adam optimizer with a learning rate of 1e-3 trained for 500 epochs.
  The attack methodology is independent of ML models used and can be evaluated easily against other architectures.
- **Model Accuracy.** The model utility is computed over the unseen test dataset across all the datasets. The test accuracy for the CENSUS dataset is 82.20%, CREDIT dataset is 77.92%, COMPAS dataset is 74.67%, and LAW dataset is 95.63%.

5.3 Metrics

We consider three main metrics for evaluating the success of attribute inference attack.

- **Precision**. The ratio of true positives to the sum of true positive and false positives. This indicates the fraction of s inferred as having a positive value by $\mathcal{A}dv$ which indeed have positive attribute value as ground truth.
- **Recall**. The ratio of true positives to the sum of true positives and false negatives. This indicates the fraction of s with positive values which are correctly inferred by $\mathcal{A}dv$.
- **F1 Score.** The harmonic mean of precision and recall computed as $2 \times \frac{precision.recall}{precision+recall}$. The highest value is one indicating perfect precision and recall while the minimum value of zero, when either precision or recall is zero.

6 TM1: EVALUATION OF ATTACK SUCCESS

We first consider **TM1**, where $\mathcal{A}dv$ has access to $\phi(x \cup s)$ and hence $\phi(x)$ and $\phi(s)$. We evaluate attack success to infer s from $\phi(x \cup s)$ (Section 6.1). Followed by this, we evaluate attack success to infer s from only $\phi(s)$ (Section 6.2).

6.1 Inferring s from $\phi(x \cup s)$

We first evaluate the simplest attack surface: $\mathcal{A}dv$ has access to the entire model explanation vector $\phi(x \cup s)$ to infer s. $\mathcal{A}dv$'s f_{adv} maps the entire explanation to s, i.e., $f_{adv} : \phi(x \cup s) \rightarrow s$. Our hypothesis is that $\phi(x \cup s)$ is distinguishable for different values of s which is captured by f_{adv}.

Table 3: TM1: Inferring s from $\phi(x \cup s)$.

Dataset	Recall	Precision	F1-Score
	Race | Sex	Race | Sex	Race | Sex
IntegratedGradients			
CENSUS	0.98 | 0.95	0.95 | 0.87	0.97 | 0.91
COMPAS	1.00 | 1.00	1.00 | 0.98	1.00 | 0.99
CREDIT	0.95 | 0.94	0.69 | 0.61	0.80 | 0.74
LAW	0.93 | 0.57	0.92 | 0.55	0.93 | 0.56
DeepLift			
CENSUS	0.99 | 0.98	0.95 | 0.93	0.97 | 0.96
COMPAS	1.00 | 0.97	1.00 | 0.98	1.00 | 0.97
CREDIT	0.95 | 0.91	0.70 | 0.61	0.81 | 0.73
LAW	0.94 | 0.88	0.92 | 0.54	0.93 | 0.67
GradientSHAP			
CENSUS	0.94 | 0.97	0.98 | 0.94	0.96 | 0.95
COMPAS	0.94 | 0.96	0.95 | 0.94	0.95 | 0.95
CREDIT	0.97 | 0.95	0.70 | 0.61	0.81 | 0.74
LAW	0.99 | 0.93	0.99 | 0.95	0.99 | 0.94
SmoothGrad			
CENSUS	0.97 | 0.95	0.94 | 0.90	0.95 | 0.93
COMPAS	0.99 | 0.99	1.00 | 0.99	0.99 | 0.99
CREDIT	0.93 | 0.92	0.70 | 0.61	0.80 | 0.73
LAW	0.99 | 0.99	1.00 | 1.00	0.99 | 0.99

As seen in Table 3, we indeed validate our hypothesis. The attack success as measured using F1-Score are high: IntegratedGradients (Sex: 0.80 ± 0.16; Race: 0.92 ± 0.07), DeepLift (Sex: 0.83 ± 0.13; Race: 0.92 ± 0.07), GradientSHAP (Sex: 0.89 ± 0.08; Race: 0.92 ± 0.06) and SmoothGrad (Sex: 0.91 ± 0.10; Race: 0.91 ± 0.10). In addition to high F1-Scores, the high precision and recall values indicate that our proposed attribute inference attack is effective to infer s from $\phi(x \cup s)$.

6.2 Inferring s from $\phi(s)$

We now consider a different setting for fine-grained analysis: can $\mathcal{A}dv$ exploit only $\phi(s)$ to infer s? Here, $\phi(s)$ is directly influenced by s while $\phi(x)$ is indirectly influence by s. Our hypothesis that $\phi(s)$ is sufficient for reasonable attack success and does not require entire model explanation $\phi(x \cup s)$ to successfully infer s. $\mathcal{A}dv$'s f_{adv} infers s using only its corresponding explanation, i.e., $f_{adv} : \phi(s) \rightarrow s$.

We validate our hypothesis as indicated by the high F1-Score: IntegratedGradients (Sex: 0.88 ± 0.09; Race: 0.88 ± 0.15), DeepLift (Sex:0.86 ± 0.09; Race: 0.87 ± 0.09), GradientSHAP (Sex: 0.78 ± 0.06; Race: 0.91 ± 0.05), and SmoothGrad (Sex: 0.79 ± 0.09; Race: 0.93 ± 0.07). In addition to high F1-Scores, the high precision and

Table 4: TM1: Inferring s from $\phi(s)$.

IntegratedGradients			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	1.00 | 0.96	0.90 | 0.75	0.94 | 0.84
COMPAS	0.99 | 0.98	1.00 | 0.94	0.99 | 0.96
CREDIT	1.00 | 1.00	0.70 | 0.82	0.61 | 0.75
LAW	0.98 | 1.00	1.00 | 1.00	0.98 | 1.00
DeepLift			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	1.00 | 1.00	0.99 | 0.70	0.99 | 0.82
COMPAS	0.94 | 1.00	0.97 | 0.81	0.95 | 0.89
CREDIT	1.00 | 0.99	0.70 | 0.61	0.82 | 0.75
LAW	0.60 | 1.00	0.99 |1.00	0.75 | 1.00
GradientSHAP			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.99 | 1.00	0.90 | 0.66	0.94 | 0.79
COMPAS	0.94 | 1.00	0.96 | 0.81	0.95 | 0.89
CREDIT	1.00 | 1.00	0.70 | 0.60	0.82 | 0.75
LAW	0.99 | 0.99	0.93 | 0.55	0.96 | 0.71
SmoothGrad			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	1.00 | 0.79	0.90 | 0.73	0.94 | 0.76
COMPAS	0.98 | 0.99	1.00 | 0.93	0.99 | 0.96
CREDIT	0.99 | 0.99	0.70 | 0.61	0.82 | 0.75
LAW	1.00 | 1.00	1.00 | 0.56	1.00 | 0.72

Table 5: TM2: Inferring s from $\phi(x)$.

IntegratedGradients			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.97 | 0.85	0.90 | 0.79	0.94 | 0.82
COMPAS	0.76 | 0.99	0.57 | 0.80	0.65 | 0.89
CREDIT	0.91 | 0.91	0.69 | 0.60	0.79 | 0.72
LAW	0.98 | 0.90	0.94 | 0.56	0.96 | 0.69
DeepLift			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.98 | 0.90	0.91 | 0.80	0.94 | 0.85
COMPAS	0.81 | 1.00	0.54 | 0.81	0.65 | 0.89
CREDIT	0.98 | 0.91	0.70 | 0.60	0.81 | 0.72
LAW	0.99 | 0.99	0.92 | 0.55	0.96 | 0.70
GradientSHAP			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.94 | 0.85	0.90 | 0.80	0.92 | 0.83
COMPAS	0.75 | 0.90	0.55 | 0.82	0.63 | 0.86
CREDIT	0.95 | 0.95	0.70 | 0.61	0.80 | 0.74
LAW	0.93 | 0.53	0.92 | 0.55	0.93 | 0.54
SmoothGrad			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.98 | 0.87	0.90 | 0.78	0.94 | 0.82
COMPAS	0.77 | 0.98	0.56 | 0.80	0.65 | 0.89
CREDIT	0.92 | 0.88	0.70 | 0.60	0.79 | 0.72
LAW	0.97 | 0.96	0.94 | 0.55	0.96 | 0.70

recall values indicate that our proposed attribute inference attack is effective to infer s from only $\phi(s)$.

Remark. *In **TM1**, the high attack success is attributed to the distinguishability of model explanations for different values of s. In other words, different values of s **explicitly** influence the model predictions as they are included in the training dataset. This in-turn results in distinguishable explanations for different values of s. This distinguishability is captured by training f_{adv} to infer s.*

7 TM2: EVALUATION OF ATTACK SUCCESS

Having shown that our proposed attack is successful in **TM1**, we now evaluate the attack success in **TM2**. We show the attack success on exploiting $\phi(x)$ (Section 7.1) followed by exploiting combination of $f_{target}(x)$ and $\phi(x)$ (Section 7.2).

7.1 Inferring s from $\phi(x)$

We evaluate the effectiveness of our attack to exploit $\phi(x)$ which are the only explanations available to $\mathcal{A}dv$. $\mathcal{A}dv$ maps $\phi(x)$ to value of s using the trained attack ML model, i.e., $f_{adv} : \phi(x) \rightarrow s$. Our hypothesis is that despite s not directly being included in the training dataset and input, some attributes in x might act as a proxy for s. Hence, s influences model predictions *indirectly* resulting in distinguishable model explanations for different values of s.

We confirm this hypothesis in Table 5 which indicates high attack success. For instance, F1-Score across four datasets for each explanation algorithm are as follows: IntegratedGradients (Sex: 0.78 ± 0.07; Race: 0.83 ± 0.12), DeepLift (Sex: 0.79 ± 0.08; Race: 0.84± 0.12), GradientSHAP (Sex: 0.74 ± 0.12 Race: 0.82 ± 0.12), and SmoothGrad (Sex: 0.78 ± 0.07; Race: 0.83 ± 0.12). Hence, censoring s is ineffective to mitigate privacy risk to attribute inference attacks.

7.2 Inferring s from $f_{target}(x) \cup \phi(x)$

Having shown the attack success on exploiting $\phi(x)$, we answer *how good are explanations + predictions combination as an attack surface for $\mathcal{A}dv$ to exploit?* We want to evaluate the impact on attack success on combining $f_{target}(x)$ with $\phi(x)$.

Given the combination $f_{target}(x) \cup \phi(x)$ as input, $\mathcal{A}dv$ trains f_{adv} to map it to s, i.e., $f_{adv} : (f_{target}(x) \cup \phi(x)) \rightarrow s$. In Table 6, we note that the attack success does not show a significant difference compared to the results in Table 5 for exploiting only model explanations. Furthermore, for Race, the attack success degrades compared to using only model explanations (Table 5). Here, we conjecture that the model predictions lower the distinguishability for f_{adv} to infer s compared to only using model explanations. These observations indicate that model explanations are a strong attack surface for $\mathcal{A}dv$ to exploit independent of model predictions.

Table 6: TM2: Inferring s from $f_{target}(x) \cup \phi(x)$.

IntegratedGradients			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.99 | 0.72	0.90 | 0.66	0.94 | 0.69
COMPAS	0.76 | 0.99	0.47 | 0.82	0.58 | 0.90
CREDIT	0.89 | 0.90	0.70 | 0.60	0.78 | 0.72
LAW	0.98 | 0.98	0.93 | 0.55	0.95 | 0.70
DeepLift			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.99 | 0.75	0.90 | 0.66	0.94 | 0.70
COMPAS	0.75 | 0.99	0.49 | 0.81	0.59 | 0.89
CREDIT	0.97 | 0.92	0.80 | 0.60	0.81 | 0.73
LAW	0.99 | 0.99	0.92 | 0.54	0.95 | 0.70
GradientSHAP			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.95 | 0.60	0.90 | 0.66	0.93 | 0.63
COMPAS	0.55 | 0.93	0.50 | 0.81	0.52 | 0.86
CREDIT	0.93 | 0.92	0.69 | 0.61	0.79 | 0.73
LAW	0.83 | 0.58	0.92 | 0.55	0.87 | 0.56
SmoothGrad			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.90 | 0.72	0.90 | 0.66	0.90 | 0.69
COMPAS	0.84 | 0.88	0.69 | 0.61	0.76 | 0.72
CREDIT	0.69 | 0.99	0.47 | 0.81	0.56 | 0.89
LAW	0.97 | 0.95	0.92 | 0.54	0.95 | 0.69

Remark. *In **TM2**, similar to **TM1**, the high attack success is attributed to the distinguishability of model explanations for different values of s. Unlike **TM1**, different values of s **implicitly** influence the model predictions via other attributes acting as proxy variables for s. This in-turn results in distinguishable explanations for different values of s which is exploited by f_{adv}.*

8 COMPARING PRIVACY RISK OF EXPLANATIONS VS. PREDICTIONS

Having shown the success of attack on model explanations, we answer *how risky are explanations compared to model predictions with respect to attribute inference attacks?* The experimental setup in our work is the same as Aalmoes et al. [1]. Hence, we report the results from Aalmoes et al. [1] as the state-of-the-art for attribute inference attacks. Specifically, we consider their PrecRec attack for both **TM1** and **TM2**.

We compare the inference capability of s from $\phi(x \cup s)$ (reported Table 3) against the inference capability of using model predictions (reported Table 7 (w/ s)). We note that the attack success in Table 3 for model explanations is higher than model predictions in Table 7 in most of the cases. Similarly, when s is not included in training data, we find that the performance reported in Table 5 is better than the results in Table 7 (w/o s).

Table 7: Reported state-of-the-art attribute inference attack success exploiting model predictions from Aalmoes et al. [1].

PrecRec Attack (w/o S)			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.91 | 0.94	0.90 | 0.69	0.90 | 0.80
COMPAS	0.97 | 0.96	0.48 | 0.82	0.64 | 0.88
LAW	0.98 | 1.00	0.95 | 0.56	0.96 | 0.72
CREDIT	0.99 | 0.97	0.69 | 0.61	0.81 | 0.75
PrecRec Attack (w/ S)			
Dataset	**Recall**	**Precision**	**F1-Score**
	Race | Sex	Race | Sex	Race | Sex
CENSUS	0.90 | 0.91	0.92 | 0.70	0.91 | 0.79
COMPAS	0.72 | 0.97	0.67 | 0.82	0.69 | 0.89
LAW	0.98 | 0.96	0.97 | 0.57	0.97 | 0.72
CREDIT	0.99 | 0.84	0.69 | 0.67	0.81 | 0.75

In summary, model explanations alone are stronger attack surface for attribute inference attack compared to model predictions.

9 RELATED WORK

We discuss some prior works which have indicated security and privacy vulnerabilities for model explanations.

Security Attacks on Model Explanations. Model explanations are sensitive to distribution shifts and adversarial examples. Model explanations do not accurately reflect the biases in ML model leading to misleading explanations which influence user trust in black box models [20]. Adversarial examples can be generated for model misclassification as well as fooling interpretations [13, 43]. The attack exploits the fact that model predictions and their interpretations are misaligned. SHAP and LIME explanations algorithms have also been shown to be vulnerable to adversarial examples [31, 33]. Counterfactual examples are an alternative approach for explanations which are not robust: they converge to different counterfactuals under a small perturbation [32]. To address these, Lakkaraju et al. [19] propose adversarial training with minimax objective to construct high fidelity explanations with respect to the worst-case adversarial perturbations. Additionally, Yeh et al. [39] propose two measures for evaluating robustness of explanations: sensitivity and infidelity, and propose algorithms to improve both.

Privacy Attacks on Model Explanations. Prior works have indicated a trade-off between transparency and privacy. Model explanations have been shown to be vulnerable to membership inference attacks where $\mathcal{A}dv$ aims to infer whether a given data record belonged to the model training data using model explanations [29]. This threat was extended to data reconstruction attacks for explanations which reveal training data instances. To incorporate membership privacy and transparency, model explanations with differential privacy have been proposed in literature [12, 26]. However, this comes at the cost of quality of explanations. Furthermore, since model explanations characterize the model's decision boundary, it can be used to steal the functionality of a model using model extraction attacks [3, 25]. None of the prior works evaluate the vulnerability to attribute inference attacks.

10 DISCUSSIONS AND CONCLUSIONS

Summary. Model explanations assign scores to attributes of an input by estimating their influence to model prediction. These model explanations potentially leak sensitive attributes. We propose the first attribute inference attack on model explanations and show their effectiveness in two threat models. We show yet another trade-off between privacy and transparency in ML models.

Attribute Privacy Risk Metric. There is a need to design data privacy risk assessment tools as required by several privacy laws such as GDPR (Article 35). However, there is limited prior work on estimating privacy risk of different sensitive attributes to inference attacks: Hannun et al. [11] propose a generic metric based on Fisher Information Loss which are shown to estimate privacy risk to attribute inference attacks. However, it is applicable only to linear and convex models and hence, not scalable to deep neural networks with non-linear and non-convex objective. Furthermore, they limit $\mathcal{A}dv$ to unbiased estimators which they indicate will be violated in the presence of $\mathcal{D}_{aux}$.

We discuss the viability of model explanations as a tool for attribute privacy risk assessment. We indicate different requirements to be satisfied for attribute privacy risk metric and indicate how model explanations satisfy them.

(1) *Independent of Attacks.* The metric should estimate the attribute privacy risk scores *without* using any specific attacks. The assigned scores should capture the root cause of attribute privacy risk, i.e., different values of s have different influence on model prediction which can be exploited by $\mathcal{A}dv$ to infer the value of s. This makes the privacy risk scores to quantify privacy risk to all possible future attacks.
   - Model explanations are independent of any specific attribute inference attacks and capture the influence of attributes to the model predictions.

(2) *Correlation with Attacks.* The attribute privacy risk scores assigned to each record's sensitive attribute should correlate with the attack success to infer s. This ensures that the privacy risk scores capture the susceptibility to attack success.
   - Model explanations can be mapped to s as shown in this work (Section 6.2) which can allow for model explanations as a relative privacy risk measure.

(3) *Efficient and Scalable.* The computation of scores should be efficient and scale to large deep neural network architectures.
   - Model explanations can be efficiently computed on deep neural networks and scalable to large models.

We leave the careful design and evaluation of attribute privacy risk metric based on model explanations for future work.

Defences Against Attribute Inference Attacks. Current literature lacks specific defences against the described attribute inference attacks as well as prior attacks leveraging model predictions. AttriGuard was proposed as a method to lower the success of $\mathcal{A}dv$'s attack ML model by adding adversarial noise to $\mathcal{A}dv$'s auxiliary data obtained from public sources [16]. This defence is more generic and can be adapted to ML models: $\mathcal{M}$ can use vulnerability of model explanations to adversarial examples, proposed in prior literature [13, 19, 20, 31–33, 39, 43], as a defence mechanism to lower the success of $\mathcal{A}dv$'s attack model. Data sanitization to remove the privacy risk while maintaining the utility of the ML model have also been explored [4]. Finallly, model explanations with differential privacy [12, 26] can possibly lower the privacy risk to attribute inference attacks as the minimize the influence of individual data records as a whole. However, using mechanisms based on pufferfish privacy [18, 36, 41] is likely address attribute inference risks. However, these have not been explored in the context of model explanations. We keep the evaluation of defences for future work.

Algorithmic Fairness and Attack Success. There are several algorithms which guarantee fairness across sensitive attributes in ML models [5, 6, 28]. It is unclear whether there is a correlation between model bias and attack success to infer s from model explanations. We speculate that since many evaluated datasets have proxy attributes to s, attribute inference attacks might still be effective (see Section 7.1). A detailed study of the impact of algorithmic fairness on attribute inference attacks of s is left for future work.

ACKNOWLEDGEMENT

The first author was supported in part by Intel (in the context of the Private-AI Institute).

REFERENCES

[1] Jan Aalmoes, Vasisht Duddu, and Antoine Boutet. 2022. Dikaios: Privacy Auditing of Algorithmic Fairness via Attribute Inference Attacks. *arXiv preprint arXiv:2202.02242* (2022).

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sy21R9JAW

[3] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. 2020. Model extraction from counterfactual explanations. arXiv:2009.01884 [cs.LG]

[4] Antoine Boutet, Carole Frindel, Sébastien Gambs, Théo Jourdan, and Rosin Claude Ngueveu. 2021. *DySan: Dynamically Sanitizing Motion Sensor Data Against Sensitive Inferences through Adversarial Networks.* Association for Computing Machinery, New York, NY, USA, 672–686. https://doi.org/10.1145/3433210.3453095

[5] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf

[6] L Elisa Celis and Vijay Keswani. 2019. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443* (2019).

[7] Amit Sharma Divyat Mahajan, Shruti Tople. 2020. Does Learning Stable Features Provide Privacy Benefits for Machine Learning Models?. In *NeurIPS PPML Workshop*.

[8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1322–1333. https://doi.org/10.1145/2810103.2813677

[9] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, San Diego, CA, 17–32. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew

[10] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. 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* (Toronto, Canada) *(CCS '18)*. Association for Computing Machinery, New York, NY, USA, 619–633. https://doi.org/10.1145/3243734.3243834

[11] Awni Hannun, Chuan Guo, and Laurens van der Maaten. 2021. Measuring Data Leakage in Machine-Learning Models with Fisher Information. In *Conference on Uncertainty in Artificial Intelligence*.

[12] Frederik Harder, Matthias Bauer, and Mijung Park. 2020. Interpretable and Differentially Private Predictions. arXiv:1906.02004 [cs.LG]

[13] Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems* 32 (2019), 2925–2936.

[14] High-Level Expert Group on AI. 2019. *Ethics guidelines for trustworthy AI.* Report. European Commission, Brussels. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[15] White House. 2020. Guidance for Regulation of Artificial Intelligence Applications. In *Memorandum For The Heads Of Executive Departments And Agencies.* https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf

[16] Jinyuan Jia and Neil Zhenqiang Gong. 2018. AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning. In *27th USENIX Security Symposium (USENIX Security 18).* USENIX Association, Baltimore, MD, 513–529. https://www.usenix.org/conference/usenixsecurity18/presentation/jia-jinyuan

[17] Emre Kazim, Danielle Mendes Thame Denny, and Adriano Koshiyama. 2021. AI auditing and impact assessment: according to the UK information commissioner's office. *AI and Ethics* (Feb 2021). https://doi.org/10.1007/s43681-021-00039-2

[18] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A Framework for Mathematical Privacy Definitions. *ACM Trans. Database Syst.* 39, 1, Article 3 (jan 2014), 36 pages. https://doi.org/10.1145/2514689

[19] Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. 2020. Robust and Stable Black Box Explanations. arXiv:2011.06169 [cs.LG]

[20] Himabindu Lakkaraju and Osbert Bastani. 2020. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20).* Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[21] European Union Law. 2018. Art. 35 GDPR Data protection impact assessment. In *General Data Protection Regulation (GDPR).* https://gdpr-info.eu/art-35-gdpr/

[22] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17).* 4768–4777.

[23] Mohammad Malekzadeh, Anastasia Borovykh, and Deniz Gündüz. 2021. Honest-but-Curious Nets: Sensitive Attributes of Private Inputs Can Be Secretly Coded into the Classifiers' Outputs. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21).*

[24] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting Unintended Feature Leakage in Collaborative Learning. *2019 IEEE Symposium on Security and Privacy (SP)* (2019), 691–706.

[25] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. 2019. Model Reconstruction from Model Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3287560.3287562

[26] Neel Patel, Reza Shokri, and Yair Zick. 2020. Model Explanations with Differential Privacy. arXiv:2006.09129 [cs.LG]

[27] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[28] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19).* Association for Computing Machinery, New York, NY, USA, 793–810. https://doi.org/10.1145/3299869.3319901

[29] Reza Shokri, Martin Strobel, and Yair Zick. 2021. On the Privacy Risks of Model Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21).* Association for Computing Machinery, New York, NY, USA, 231–241. https://doi.org/10.1145/3461702.3462533

[30] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17).* 3145–3153.

[31] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. arXiv:1911.02508 [cs.LG]

[32] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual Explanations Can Be Manipulated. arXiv:2106.02666 [cs.LG]

[33] Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Feature Attributions and Counterfactual Explanations Can Be Manipulated. arXiv:2106.12563 [cs.LG]

[34] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. *ArXiv* abs/1706.03825 (2017).

[35] Congzheng Song and Vitaly Shmatikov. 2020. Overlearning Reveals Sensitive Attributes. In *International Conference on Learning Representations.* https://openreview.net/forum?id=SJeNz04tDS

[36] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. 2017. Pufferfish Privacy Mechanisms for Correlated Data. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '17).* Association for Computing Machinery, New York, NY, USA, 1291–1306. https://doi.org/10.1145/3035918.3064025

[37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17).* 3319–3328.

[38] Elham Tabassi, Kevin J. Burns, M. Hadjimichael, Andres Molina-Markham, and Julian Sexton. 2019. A Taxonomy and Terminology of Adversarial Machine Learning. In *NIST Interagency/Internal Report.* https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

[39] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems* 32 (2019), 10967–10978.

[40] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF).* 268–282. https://doi.org/10.1109/CSF.2018.00027

[41] Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. 2022. Attribute privacy: Framework and mechanisms. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* 757–766.

[42] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. 2021. Leakage of Dataset Properties in Multi-Party Machine Learning. In *30th USENIX Security Symposium (USENIX Security 21).* USENIX Association, 2687–2704. https://www.usenix.org/conference/usenixsecurity21/presentation/zhang-wanrong

[43] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable Deep Learning under Fire. In *29th USENIX Security Symposium (USENIX Security 20).* USENIX Association, 1659–1676. https://www.usenix.org/conference/usenixsecurity20/presentation/zhang-xinyang