Dikaios: Privacy Auditing of Algorithmic Fairness via Attribute Inference Attacks

Jan Aalmoes
Univ Lyon, INSA Lyon, Inria, CITI
Lyon, France
jan.aalmoes@inria.fr

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
Machine learning (ML) models have been deployed for high-stakes applications (e.g., criminal justice system). Due to class imbalance in the sensitive attribute observed in the datasets, ML models are unfair on minority subgroups identified by a sensitive attribute, such as Race and Sex. Fairness algorithms, specially in-processing algorithms, ensure model predictions are independent of sensitive attribute for fair classification across different subgroups (e.g., male and female; white and non-white). Furthermore, ML models are vulnerable to attribute inference attacks where an adversary can identify the values of sensitive attribute by exploiting their distinguishable model predictions. Despite privacy and fairness being important pillars of trustworthy ML, the privacy risk introduced by fairness algorithms with respect to attribute leakage has not been studied. In addition to different fairness metrics, we identify attribute inference attacks as an effective measure for auditing blackbox fairness algorithms to enable model builder to account for privacy and fairness in the model design. More precisely, we proposed Dikaios, a privacy auditing tool for fairness algorithms which leveraged a new effective attribute inference attack that account for the class imbalance in sensitive attributes through an adaptive prediction threshold. Dikaios can be used by model builders to estimate the attribute privacy risks of their model with or without the sensitive attribute in model training. We exhaustively evaluated Dikaios to perform a privacy audit of two in-processing group fairness algorithms (i.e., reductions and adversarial debiasing) over five datasets. First, we show that our attribute inference attack with adaptive prediction threshold significantly outperforms prior attacks, and second, we highlighted the limitations of in-processing fairness algorithms to ensure indistinguishable predictions across different values of sensitive attributes. Indeed, the attribute privacy risk of these in-processing fairness schemes is highly variable according to the proportion of the sensitive attributes in the dataset. This unpredictable effect of fairness mechanisms on the attribute privacy risk can be an important limitation on their utilization which has to be accounted by the model builder.

KEYWORDS
Fairness, Data Privacy, Attribute Inference Attacks, Deep Learning.

1 INTRODUCTION
Machine learning (ML) are extensively deployed for high-stakes decision making applications such as criminal justice, finance and healthcare. Model builders design and train ML models for deployment as a service to the Cloud, referred to as Machine Learning as a Service (MLaaS). Such applications require the design of trustworthy ML models which have been stressed by government organizations including European Union’s AI guidelines1 as well as companies such as Google2. Two of the important aspects of trustworthy ML are (a) fairness across different subgroups and (b) privacy of an individual’s personal and potentially sensitive data.

The class imbalance in sensitive attributes, such as Sex and Race, of the dataset typically results in a distinguishable model performance across minority and majority subgroups where the model predictions are skewed for a particular subgroup. This results in a discriminative behaviour of the models which makes fairness in ML crucial, specially, in critical applications such as criminal justice, healthcare and admission selection [30]. Algorithmic fairness addresses the problem of difference in model performance over different subgroups identified by a particular sensitive attribute, such as Sex and Race [5, 19, 48]. For instance, the goal of in-processing group fairness algorithms is to add constraints during the model training to ensure indistinguishable model predictions over different subgroups identified by sensitive attributes (e.g., males vs. females, whites vs. non-whites).

On the other hand, ML models, when trained on sensitive data, are vulnerable to inference attacks which violate the privacy of an individual contributing their data [14, 15, 40]. Specifically, an adversary can exploit the distinguishable model predictions to infer whether an individual’s data record was part of the model’s training [6] or to infer a sensitive attribute of an individual’s data record. Although, Differential Privacy (DP) is defined as a defence to protect a data record’s training data membership status, however, it is not effective to protect against population level attacks due to dataset skewness of sensitive attributes [8]. Consequently, even under DP, properties of datasets and sensitive attributes can be inferred accurately [21].

In this work, we leverage the relation between fairness and attribute privacy. Indeed, group fairness ensures the model predictions to be independent of the sensitive attribute, i.e., indistinguishable model predictions for different subgroups. This preserves privacy against attribute inference attack which exploits the distinguishability of model predictions to infer sensitive attributes. Consequently, attribute inference attacks could be used as a privacy auditing tool to evaluate the privacy risks of fairness algorithms with respect to sensitive attributes.

Unfortunately, existing attribute inference attacks cannot be used for privacy auditing of fairness algorithms. Firstly, the threat

---

1. EU guidelines on ethics in AI[35]
2. https://ai.google/responsibilities/responsible-ai-practices
model considered in prior work are not practical wherein they assume a strong adversary with access to non-sensitive features and the corresponding model predictions which are used to infer the sensitive attribute. Hence, the adversary cannot choose the inputs to query the model as it contains the sensitive attribute that the adversary aims to infer [14, 15, 47]. Secondly, they assume that the sensitive attribute is used to train the model and does not extend to the setting where model builder censors the sensitive attribute by removing it from the training dataset for privacy [14, 15, 47]. Finally, they do not account for class imbalance in sensitive attributes which strongly influences the attack success [11, 41].

In this paper, we deeply analyse the impact of fairness algorithms on the privacy risks through the lens of attribute inference attacks. We design a new attribute inference attacks by accounting for the class imbalance of sensitive attributes in training datasets using an adaptive threshold over the adversary’s attack model output predictions to increase the adversary’s success. The attack is efficient even when the model is trained without the sensitive attributes (censoring for privacy by model builder) which makes them practical. Finally, the attack, adapted for model builder by including the sensitive attribute to train the model, forms the basis of Dikaios privacy auditing tool for fairness algorithms to estimate worst-case attribute privacy risks.

We claim the following main contributions:

1. **Propose a new attribute inference attack.** The attacks account for the imbalance of sensitive attributes in datasets by computing an adaptive prediction threshold to improve adversary’s attack success. We consider two variants of the attack according to the threshold optimized for a specific attack metric: AUC-ROC and Precision-Recall. (Section 4)

2. **Propose Dikaios, the first attribute privacy auditing tool for model builders to estimate the attribute privacy risk.** by leveraging our new inference attack to measure the distinguishability of model predictions of different subgroups. (Section 5)

3. **Evaluate our inference attack in a practical threat model** where the sensitive attribute is included in training data or censored by the model builder for privacy. The best attack outperforms state-of-the-art inference attacks up to 20% on average across five benchmark datasets. (Section 7)

4. **Estimate attribute privacy risks for in-processing group fairness algorithms using Dikaios.** We find that attribute privacy risk of in-processing fairness schemes is highly variable on changing the proportion of the sensitive attributes. This unpredictable effect of fairness algorithms on the privacy cost is an important limitation on their utilization to be considered by the model builder. (Section 8)

2 BACKGROUND

We present a background on ML and notations (Section 2.1), definitions for attribute inference attacks (Section 2.2) and group fairness (Section 2.3).

2.1 Machine Learning

ML algorithms learn a function \( f_{\text{target}} : X \rightarrow Y \), where \( X \) is the set of all possible non-sensitive input features and \( Y \) is the space of corresponding labels. The model parameters \( \theta \) are updated iteratively using the gradient of the loss between model’s prediction and the true label. The model training is done using training dataset \( D \subset X \times Y \). The function \( f_{\text{target}} \) converges to a minima where it successfully maps the inputs to the corresponding labels. In addition to non-sensitive input features \( X \) and corresponding labels \( Y \) we denote the set of all possible sensitive attributes as \( S \). We consider a data record \( z = (x, s, y) \) with non-sensitive features \( x \), sensitive attribute \( s \) where \( (x, s) \in X \times S \), and classification label \( y \). Note that we assume both \( y \) and \( s \) are binary variables.

2.2 Attribute Inference Attacks

ML models leak sensitive attributes in the training data by exploiting the distinguishable output predictions corresponding to different values of sensitive attributes. For instance, the output prediction distribution for “Males” is different from the output prediction distribution for “Females” (see Figure 2).

Prior attribute inference attacks have a setting similar to data imputation where the goal is to infer the missing sensitive attribute given other non-sensitive features and output predictions [11, 14, 15, 41]. For instance, Fredrikson et al. [14, 15] consider the setting (a) where the adversary (\( \mathcal{A} \)) aims to infer \( s \) given \( x \) and \( f_{\text{target}}((x, s)) \). Yeom et al. [47] improve Fredrikson et al.’s attack by assuming a distribution over sensitive attributes \( P_{s} \) to infer the value of \( s \) for a particular data record. These attacks assume the model was trained on the input \((x, s)\). Song et al. [41] proposed inferring sensitive attributes using an intermediate representation which can be extended to output predictions. Mahajan et al. [11] indicate an attribute inference attack using output predictions within the context of domain generalization. Both Song et al. and Mahajan et al. use an ML model to infer \( s \) given \( f_{\text{target}}(x) \). They use the default threshold of 0.5 over the output prediction probabilities for the attack model to classify \( s \leftarrow 0 \) or \( s \leftarrow 1 \).

We differentiate attribute inference attacks from property inference attacks where instead of learning about the sensitive attributes of individual data records, \( \mathcal{A} \) aims to infer a global statistical properties of the dataset [17, 31, 51, 51]. For instance, inferring the ratio of some property (number of males to number of females) in the dataset used to train the model. We focus on attribute inference attacks and not property inference attacks where we aim to infer the specific value of sensitive attribute of a data record instead of a global dataset statistics.

2.3 Algorithmic Fairness

Two notions are used for fairness in ML: group fairness and individual fairness. While group fairness aims to add constraints to model training such that different subgroups (categorized by specific \( s \) such as \( \text{Sex} \) and \( \text{Race} \)) are treated equally, individual fairness ensures that two data records with same features except for value of \( s \) have the same model prediction. In this work, we focus on the notion of group fairness where the model prediction is independent of the sensitive attributes and hence indistinguishable for different values of \( s \) [50]. This notion of indistinguishability of output predictions for sensitive attributes can be formalized for a set of inputs \( X \) a, a set of sensitive attributes \( S \) and a set of true labels \( Y \), by different fairness definitions:
Demographic Parity. Prediction $f_{\text{target}}(X)$ satisfies demographic parity if the prediction $f_{\text{target}}(X)$ is independent of $S$. Hence, $P(f_{\text{target}}(X) = \hat{y})$ is the indistinguishable for all values of sensitive attributes $S$ given as $P(f_{\text{target}}(X) = \hat{y}) = P(f_{\text{target}}(X) = \hat{y}|S = s)$.

Equality of Odds. Prediction $f_{\text{target}}(X)$ satisfies equalized odds if $f_{\text{target}}(x)$ is conditionally independent of $S$ given $Y$.

Equality of Opportunity. Prediction $f_{\text{target}}(X)$ satisfies equality of opportunity with regards to specific class $y$ if the prediction $f_{\text{target}}(x)$ is independent to $S$ conditioned on $Y = y$.

Where $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and $S : \Omega \rightarrow S$ are random variables.

For both equality of opportunity and equality of odds, for a particular true label $y$, the output prediction probability $P(f_{\text{target}}(X) = \hat{y})$ is indistinguishible for all values of $s$: $P(f_{\text{target}}(X) = \hat{y}|Y = y) = P(f_{\text{target}}(X) = \hat{y}|S = s, Y = y)$.

The above definitions can be achieved using three main fairness mechanisms: (a) pre-processing, (b) in-processing and (c) post-processing. Pre-processing algorithms such as reweighing requires access to the train data and assigns weights to the data records to remove discrimination [24]. In-processing fairness algorithms such as adversarial debiasing [50] and reductions [2] add constraint during the model training to ensure fairness. Post-processing techniques, in turn, hide the bias in output predictions to satisfy the above fairness constraints but the underlying model is still biased.

To evaluate how far a model is from those definitions, we use the disparate impact metric: the ratio of the accuracy for the unprivileged class to the privileged class. Formally, this can be written as $\frac{\text{Pr}(f_{\text{target}}(x)|s=\text{unprivileged})}{\text{Pr}(f_{\text{target}}(x)|s=\text{privileged})}$, where the unprivileged class is in minority (e.g., “non-white” or “female”) whereas privileged is in majority class (e.g., “white” or “male”). A disparate impact of 1 means that the model is fair and the typical industry standard for a fair model is that ratio should be 80% (i.e., 0.8).

### 3 PROBLEM STATEMENT

Fairness algorithms claim to reduce the distinguishability of different values of $s$ [2, 5, 24]. However, these algorithms do not result in uniform fairness level obtained on varying the proportion of $s$. Hence, a fair model might still allow an adversary to infer $s$ using distinguishability in output predictions, leaking $s$ and violating attribute privacy [53]. Hence, this requires model builder ($M$) to understand when to use fairness techniques to effectively mitigate bias without leaking $s$. We identify attribute inference attacks as an effective measure for auditing blackbox fairness algorithms and describe the intuition below.

Relation between fairness and attribute privacy. The notion of attribute inference attacks (privacy) is directly related to the notion of group fairness. The group fairness aims to minimize distinguishability of model predictions conditioned on different values of $s$.

The loss in attribute privacy is estimated using attribute inference attack which infer $s$ by exploiting the distinguishability in model predictions. This allows us to use attribute inference attacks as a privacy auditing tool which is essential for $M$ to (a) estimate attribute privacy risks of fairness algorithms and (b) analyze the effectiveness of fairness mechanisms.

### Limitations of prior work.

We first describe the limitations of prior attribute inference attacks to understand why they cannot be used for privacy auditing of fairness algorithms.

1. **Not a practical threat model**: Some attacks assume a strong $\mathcal{A}$ with access to $x$ and corresponding $f_{\text{target}}((x, s))$. $\mathcal{A}$ trains an attack ML model to infer the $s$. Furthermore, all these attacks consider $s$ in the training dataset, i.e., $\mathcal{A}$ only sees the $f_{\text{target}}((x, s))$ and not the input data record (since they have the $s$ in the features).

2. **Do not account for class imbalance in $s$**: Model predictions are strongly skewed towards particular value of $s$ (typically males for Sex or whites for Race). This is due to the class imbalance in $s$ of the datasets which strongly influences the attack success. This has not been accounted for in the prior work.

3. **Do not account for censoring of $s$**: $M$, with the goal to protect the $s$, can censor the $s$ by removing it from the dataset for training model. This setting of censoring $s$ has been explored in the closely related property inference attack [51] but not within the context of attribute inference attacks. The current attacks are not applicable to this threat model.

In the view of the above limitations of prior work, our goal is to design practical blackbox attribute inference attacks for privacy auditing of fairness algorithms to estimate the leakage of $s$ from model predictions.

### Research Questions.

To sum up, we identify and address the following main research questions:

- **RQ1**: How can we design practical blackbox attribute inference attack as an effective tool to audit fairness algorithms?
- **RQ2**: How effective are fairness algorithms to reduce distinguishability between different values of $s$?
- **RQ3**: What are the privacy risks for in-processing fairness algorithms with respect to attribute inference attacks?

### 4 DESIGNING ATTRIBUTE INFERENCE ATTACKS

Having discussed the limitations of prior attribute inference attacks for privacy auditing of algorithmic fairness (Section 3), we describe the design requirements for proposing a new attribute inference attack.

#### Requirements for Attack Design.

To address limitation **L1**, we should consider a practical threat model which allows $\mathcal{A}$ to choose the input (without $s$) and use only the corresponding predictions $f_{\text{target}}(x)$ to infer $s$. For limitation **L3**, the attack should ideally be designed for setting with and without $s$ in training data. For addressing limitation **L2** for privacy auditing of fairness algorithms, which are specifically trained on highly imbalanced datasets, attacks are required to account for class imbalance of $s$.

We first describe the threat model considered for the proposed attack (Section 4.1) followed by the description of the attacks (Section 4.2).
4.1 Threat Model: A’s Perspective

Here, we describe the A’s knowledge and capabilities for attribute inference attacks. We consider a blackbox A with no knowledge about the model’s parameters or architecture. A has an API access (similar to setting seen in MLaaS) to pass an input and output the corresponding output prediction. The target model $f_{\text{target}}$ is trained on training data $D$. Unlike prior work on attribute inference attacks [14, 15, 47], we consider the practical case where $s$ is not part of the input (referred as “w/o $s$”). In other words, the model was trained on non-sensitive input features $x$. This is the worst case scenario for A where $s$ is censored by model builder for privacy reasons.

Additionally, we assume that A has access to a non-overlapping auxiliary dataset $D_{aux}$ with inputs $(x', s', y')$. We assume $D_{aux}$ is sampled from the same distribution as $D$.

The goal of A is to infer $s$ despite the main classification task of $f_{\text{target}}$ is prediction of $y$ using $f_{\text{target}}(x)$. Here, A trains an attack model $f_{\text{adv}} : f_{\text{target}}(x) \rightarrow s$. $f_{\text{adv}}$ is trained on the dataset of target model predictions $f_{\text{target}}(x')$ from $D_{aux}$ with $s'$ as the target label. By default, $f_{\text{adv}}(f_{\text{target}}(x))$ indicates the probability scores across different class values of $s$ with the threshold as 0.5. In other words, if $f_{\text{adv}}(f_{\text{target}}(x)) > 0.5$ the predicted sensitive attribute $s \leftarrow 1$ else $s \leftarrow 0$.

4.2 Adaptive Threshold

For fairness algorithms, the datasets are highly imbalanced over $s$ (see Table 1). Hence, using the default classification threshold for the attack model of 0.5 could result in poor attack performance (close to random guess as shown in Section 7.2).

We propose an attribute inference attacks with adaptive thresholds for $f_{\text{adv}}$ to account for imbalance in $s$. The class imbalance in attributes is accounted by moving the default classification threshold for attack model from 0.5 to optimal value. A computes the optimal classification threshold $\tau^*$ for $f_{\text{adv}}$ on the auxiliary dataset which is then used to infer $s$ on any arbitrary target record.

We consider two variants of the attack where each variant computes the optimal threshold by optimizing for a specific attack metric: AUC-ROC (AucRoc) and Precision-Recall (PrecRec). The attack aims to increase the success of A by finding the optimal threshold over the model predictions to estimate the correct sensitive attribute. We change the objective of A, i.e., instead of finding a threshold that maximizes the attack accuracy on the auxiliary dataset, we proposed to find thresholds that maximize the power of A for different attack success metrics (e.g., precision, recall and F1-Score). Different objectives will result in different thresholds, which ultimately result in different results.

AucRoc: Calibration using AUCROC curves

The attack based on the AucRoc objective computes the optimal threshold using the AUCROC curves which plot the set of probabilistic predictions (also referred as “soft labels”) on the evaluation dataset. The choice of the threshold made by A determines the true positive (TPR) and false positive rates (FPR) rate on the minority class. The optimal threshold for a theoretically perfect attack would result in zero FPR and perfect results on TPR. A’s goal is to approach this optimal value of TPR and FPR. Formally, we define AUCROC curve as a function which maps each threshold value in the set $T$ to the corresponding FPR and TPR values given as $\text{roc} : T \rightarrow [0,1]^2$. Hence, the best threshold $\tau^*$ for a theoretically perfect attack gives $\text{roc}(\tau^*) = (0,1)$.

The attack based on AucRoc computes $\tau^* = \arg \min_{\tau \in T} \text{roc}(\tau)$ where

$$\text{roc}(\tau) : \{\text{fpr}, \text{tpr}\} \rightarrow [0,1]^2$$

In other words, we can efficiently compute $\tau^*$ using $\text{roc}$ (referred as “Youden’s J statistic”) as difference between TPR and FPR across all the thresholds. The optimal threshold chosen based on the highest $J$ statistic. Hence in the attack based on AucRoc, the power of A increases by trying to increase TPR - FPR. Finding the threshold using AucRoc attack is better when there are roughly equal numbers of values for each class of s. This is because the AUCROC curves provide an optimistic view of class imbalance which may not give the best results [10].

PrecRec: Calibration using Precision-Recall Curves

We provide an alternative for computing $\tau^*$ when there is a moderate to large class imbalance. This can be effectively done using the precision-recall curves which forms the basis of the variant of the attack based on the PrecRec objective. Unlike the AUCROC curves, precision-recall curves capture the trade-off between TPR and the positive predictive value (precision) for a predictive model across different probability thresholds.

The precision-recall curve is generated by computing the precision and recall with the true labels after thresholding the prediction probabilities by the target model across all thresholds. A random guess for precision-recall curve is the horizontal line with the precision value computed over the positive class examples in the dataset (shown in Appendix Figure 3). We define Precision-Recall curve as a function which maps each threshold value in the set $T$ to the corresponding Precision and Recall values. Hence, the best threshold $\tau^*$ for a theoretically perfect attack gives $\text{roc}(\tau^*) = (1,1)$ where the values of both perfect precision and recall values resulting in perfect F1-score. $\tau^*$ is the threshold corresponding to the maximum F1-Scores across all thresholds.

5 DIKAIOS: PRIVACY AUDITING BY M

Having described the novel attribute inference attack proposed in the work, now we adapt it to a different threat model for using it as a privacy auditing tool, Dikaios, by M. We describe the M’s objective, knowledge and capabilities (Section 5.1) and some of the properties satisfied by Dikaios (Section 5.2).

5.1 Threat Model: M’s Perspective

The threat model described in Section 4.1 is from A’s perspective for a practical blackbox adversary with minimal knowledge to infer $s$. However, M’s perspective and threat model is different. M aims to assess the worst-case attribute privacy risk of different fairness algorithms and has additional training data knowledge which is lacking for A.

Similar to A threat model in Section 4.1, we assume that M has access to a non-overlapping auxiliary dataset $D_{aux}$ with inputs $(x', s', y')$, sampled from the same distribution as $D$. Despite having access to $D$, assuming access to $D_{aux}$ keeps the threat model in blackbox setting.
Additionally, \( M \) has access to the sensitive attribute which is included in the training dataset of \( f_{\text{target}} \). The input is hence \((x, s)\) which we refer as "w/ s". This helps estimating the worst case risk for the \( M \). The main difference of including \( s \) to train the model results in a direct correlation of output predictions to \( s \) which is not the case for \( \mathcal{A} \) threat model after censoring.

Given the target model’s output prediction \( f_{\text{target}}((x, s)) \), \( M \) trains an attack model \( f_{\text{mb}} : f_{\text{target}}(X \times S) \rightarrow S \). \( f_{\text{mb}} \) is trained to map \( f_{\text{target}}((x', s')) \) to the target label \( s' \) using \( D_{\text{aux}} \). Here, the decision of estimating \( s' \) from \( f_{\text{mb}}(f_{\text{target}}((x, s))) \) uses the adaptive thresholds computed as part of the attack described in Section 4.2.

5.2 Summary of Dikaios Properties
We summarize some of the properties of Dikaios which make it a practical tool for privacy auditing against attribute inference.

P1 Practical blackbox threat model: An ideal blackbox \( \mathcal{A} \) sends an input to the model and observes the corresponding predictions. The output prediction is the only information available to perform the attack to \( \mathcal{A} \).

P2 Accounting for class imbalance in \( s \): \( f_{\text{mb}}(f_{\text{target}}(X \times S)) \) is highly skewed towards a particular class of \( s \) which impacts the performance of attribute inference attacks. The attacks compute \( \sigma' \) over \( f_{\text{mb}}(f_{\text{target}}(X \times S)) \) to estimate \( \hat{s} \).

P3 Worst-case attribute privacy risk: The attack is operable in the threat model where \( M \) by including the \( s \) (available to \( M \)) in the training dataset. This results in a direct correlation between \( f_{\text{target}}((x, s)) \) and \( s \) to improve the attack success.

In this work, we demonstrate against the specific application of algorithmic fairness. The attacks while specifically tailored to algorithmic fairness, result in state-of-the-art performance. Hence, Dikaios is generally applicable for auditing attribute privacy risks in ML models in other applications too.

6 EXPERIMENT SETUP
We now describe our experimental setup: description of the datasets (Section 6.1), evaluation metrics (Section 6.2), baseline approaches to compare against proposed attacks (Section 6.3), and the fairness algorithms considered in this work for privacy auditing (Section 6.4).

6.1 Datasets
We consider five real world datasets, including criminal justice, income prediction, health dataset, college performance and credit applications, to illustrate the effectiveness of the proposed attribute inference attacks and Dikaios for privacy auditing.

US adult income dataset (CENSUS) comprises of 48,842 data records with 95 attributes about individuals from 1994 US Census data. The attributes include marital status, education, occupation, job hours per week among others. The classification task is to estimate whether an individual makes an income of 50k per annum. We use 21,658 data records for \( \mathcal{D} \) and 9,282 data records for testing dataset. We use the 9,282 test data records as \( D_{\text{aux}} \) for training the attack model and evaluate it on 9,282 data records from \( \mathcal{D} \).

Recidivism dataset (COMPAS) is used for commercial algorithms by judges and parole officers for estimating the likelihood of a criminal re-offending. The classification task is binary: whether a criminal will re-offend or not, and contains around 10,000 criminal defendants in Florida. The dataset contains 7 attributes and we use 4,320 data records for \( \mathcal{D} \) and 1,852 data records for testing dataset. We use the 1,852 test data records as \( D_{\text{aux}} \) for training the attack model and evaluate it on 1,852 data records from \( \mathcal{D} \).

Medical Expenditure Panel Survey dataset (MEPS) contains around 15,830 records of different patients using medical services by capturing the trips made to clinics and hospitals. The classification task is to predict the utilization of medical resources as ‘High’ if the sum of the number of office based visits, outpatient visits, ER visits, inpatient nights and home health visits, is greater than 10. We use 11,081 data records for \( \mathcal{D} \) and 4,749 training data records for testing dataset. We use the 4,749 test data records as \( D_{\text{aux}} \) for training the attack model and evaluate it on 4,749 data records from \( \mathcal{D} \).

Law school dataset (LAW) 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 data is collected based on a survey conducted by Law School Admission Council across 163 law schools in the United States. Given this data, a school predicts if an applicant will have a high first year average grade. The dataset has four attributes and we use 14,322 data records for training and 6,138 data records for testing. We use the 6,138 test data records as \( D_{\text{aux}} \) for training the attack model and evaluate it on 6,138 data records from \( \mathcal{D} \).

UCI Credit Card dataset (CREDIT) is from the UCI Machine Learning dataset repository and contains information about different credit card applicants. The classification task is binary indicating whether the application was approved or not. The dataset contains 30,000 records with 24 attributes for each record. We use 21,000 train data records and 9,000 test data records. We use the 9,000 test data records as \( D_{\text{aux}} \) for training the attack model and evaluate it on 9,000 data records from \( \mathcal{D} \).

In all the datasets, we consider the Sex ("Male and Female") and Race ("White and Non-White") as the sensitive binary attributes. For CREDIT dataset, we use the Age attribute classified as "Young" or "Old" as sensitive. For brevity, we refer Age attribute as Race in evaluation tables. We summarize the class imbalance for different datasets in Table 1.

### Table 1: Summary of class imbalance in \( s \) for the datasets.

| Dataset | SEX | Race       |
|---------|-----|------------|
|         | Male | Female    | White | Non-White |
| CENSUS  | 68%  | 32%       | 90%   | 10%       |
| COMPAS  | 81%  | 19%       | 49%   | 51%       |
| LAW     | 57%  | 43%       | 94%   | 6%        |
| MEPS    | 53%  | 47%       | 36%   | 64%       |
| CREDIT  | 61%  | 39%       | 70%   | 30%       |

Target model accuracy on primary task. We train the ML models on primary classification tasks for all the five datasets. We found the baseline accuracy for COMPAS dataset with 74.89%, CENSUS dataset with 82.95% accuracy, 86.31% accuracy for MEPS dataset, and 95% for LAW dataset and 81% for CREDIT dataset.
### 6.2 Metrics
We consider the following metrics for evaluating the success of attribute inference attacks.

**Precision** is the ratio of true positives to the sum of true positive and false positives. This indicates the fraction of $s$ inferred as the positive class by $\mathcal{A}$ which indeed belong to positive class.

**Recall** is the ratio of true positives to the sum of true positives and false negatives. It indicates the fraction of $s$ predicted as positive class which are correctly inferred as positive class by $\mathcal{A}$.

**F1-Scores** is the harmonic mean of precision and recall and computed as $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$.

**Balanced Accuracy** is the average of the proportion of correct predictions of each class of $s$ individually. This is used to account for imbalance in the classes. We refer to “Balanced Accuracy” as “Accuracy” in the rest of the paper for brevity. Note that using naive accuracy as an attack metric is misleading when the datasets have significant class imbalance (see Table 1).

### 6.3 Prior Attribute Inference Attacks
We use three main prior attribute inference attacks as baselines with different assumptions for $\mathcal{A}$’s knowledge. We describe these attacks below.

Fredrikson et al. [14, 15] attack assumes that $\mathcal{A}$ has access to $\text{fi}_{\text{target}}((x, s))$ from the target classifier $\text{fi}_{\text{target}}$ and $x$. $\mathcal{A}$ trains $f_{\text{adv}}$ to infer $s$, i.e., $f_{\text{adv}} : \text{fi}_{\text{target}}(X \times S) \times X \rightarrow S$. However, this is a strong assumption for a blackbox setting.

Yeom et al. [47] assume a distribution $P_s$ over $s$ which is used to estimate the value of $s$ for an arbitrary data record. They propose four different variants of attribute inference attacks based on assumptions on $P_s$. Attack 1 leverages membership oracle to determine the value of $s$. Attack 2, 3 and 4 assume different distributions over $s$ for $\mathcal{D}$.

1. **Attack 1: Membership Inference Oracle.** The attack predicts the value of $s$ for a given data record based on the output of membership inference oracle, which predicts whether a given data record was used in the model’s training.

2. **Attack 2: Assuming Normal distribution.** The attack assumes a distribution over the prediction error $\epsilon = \text{fi}_{\text{target}}(D_{\text{aux}}) - y$ where $\text{fi}_{\text{target}}$ is trained on the dataset $D_{\text{target}}$. $D_{\text{aux}}$ is $\mathcal{A}$’s auxiliary data with $y$ as the true labels. The attack assumes that the error $\epsilon$ follows multivariate normal distribution for which the mean and standard deviation are computed as $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i$ and $\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\epsilon_i - \hat{\mu})(\epsilon_i - \hat{\mu})^T$. This mean and standard deviation are computed by $\mathcal{A}$ using $D_{\text{aux}}$ which is used to estimate the unknown $s$ for arbitrary data records.

3. **Attack 3: Assuming uniform distribution.** Here, the probability function is described as a uniform distribution instead of normal distribution. Let $n$ be the index of $s$ as defined by $\text{fi}_{\text{target}}$ architecture and $S = \{s_1, \cdots, s_m\}$ the set of possible values for $s$. The inferred value of $s$ is described as $\hat{s} = \arg \max_{s \in S} P(s = s | f_{\text{target}}(x) = y, \hat{\Sigma})$ where $f_{\text{target}}$ is the distribution of $s$ in $S$. The error of $\text{fi}_{\text{target}}$ is $\epsilon = \epsilon(s', x, y) = y - \text{fi}_{\text{target}}(x(s'))$ where $x(s') = (x_1, \cdots, x_{n-1}, s', x_{n+1}, \cdots)$.

Song et al. [41]/Mahajan et al. [11] has a practical threat model compared to above attacks and assumes that the input does not have $s$. $\mathcal{A}$ only observes $\text{fi}_{\text{target}}(x)$ for a given $x$. $\mathcal{A}$ trains $f_{\text{adv}}$ on $D_{\text{aux}}$ to map the output predictions from $\text{fi}_{\text{target}}$ to $s$, i.e., $f_{\text{adv}} : \text{fi}_{\text{target}}(x') \rightarrow s'$. However, the classification threshold is default value of 0.5 for inferring $s$ from $f_{\text{adv}}$, i.e., attack model predicts $s$ as “1” if output probability $>0.5$ and class “0” otherwise. A threshold of 0.5 is acceptable when the number of records in each class in $s$ are same. However, in most fairness applications, the majority class dominates over the minority class which shifts the attack model’s threshold over prediction probability. This makes this attribute inference attack ineffective to audit fairness algorithms.

**Qualitative comparison with prior attacks.** We summarize the attack setup of prior work along with our proposed attacks in Table 2. We indicate the attack features used for performing the attack and omit $\mathcal{A}$’s knowledge of $D_{\text{aux}}$ to train an attack model which are used by all the attacks.

| Literature        | Features | $s \in \mathcal{D}$ | Imbalance in $s$?
|-------------------|----------|----------------------|------------------|
| Fredrikson et al. [14, 15] | $x. \, \text{fi}_{\text{target}}()$ | ✓ | × |
| Yeom et al. [47]   | $x, y. \, \text{fi}_{\text{target}}(), P_s$ | ✓ | × |
| Song et al. [41]   | $\text{fi}_{\text{target}}()$ | ✓ | × |
| Mahajan et al. [11] | $\text{fi}_{\text{target}}()$ | ✓ | ✓ |
| **Our Work**       | $\text{fi}_{\text{target}}()$ | ✓ | ✓ |

Fredrikson et al. [14, 15] and Yeom et al. [47] assume a strong threat model (see limitation L1). While Song et al. [41] and Mahajan et al. [11] while having a practical threat model, they do not account for class imbalance in $s$ (see limitation L2). None of the prior attacks consider the case where $s$ is not included in the dataset (see limitation L3).

Compared to prior work, DIKAIOS works even when $s$ is not included in $\mathcal{D}$ of $\text{fi}_{\text{target}}$. This a more practical threat model. Further, we account for the class imbalance in $s$ by using adaptive threshold over the attack model prediction to increase the power of adv. Hence, DIKAIOS can analyze fairness algorithms which is not possible for other attribute inference attacks.

### 6.4 In-Processing Fairness Algorithms
We demonstrate the attacks on in-processing algorithms which aim to correct the bias of the model by adding fairness constraints during training. However, the proposed attacks are independent of specific fairness algorithms used and hence, applicable to other pre-processing and post-processing fairness algorithms. We specifically consider two algorithms: Exponentiated gradient reduction (REductions) [2] and Adversarial Debiasing (AdvDebias) [27, 50], as described below.

**Reductions** [2] enforces group fairness by the training $\text{fi}_{\text{target}}$ using a set of linear constraints thus reducing the possible set of value for parameters $\theta$. Those constraints are ensure equality such that $E \left[ \text{fi}_{\text{target}}(X) | s = s_i, y = y_j \right] = E \left[ \text{fi}_{\text{target}}(X) | y = y_j \right]$ $\forall (s_i, y_j) \in \{0, 1\}^2$. Computationally, the equality is replaced with
inequalities to allow for approximation with some error with true solution. The resulting problem admits a solution in the sense that it becomes a saddle point problem when introducing Lagrange multiplier. **REDUCTIONS** obtains the solution using the Freund-Schapire numerical scheme [16].

**AdvDebias** [27, 50] trains \( f_{target} \) alternatively to solve a zero sum game with a another ML “inference” model \( f_{inf} \). \( f_{inf} \) is trained to differentiate between the predictions \( P(y|x, s=0) \) and \( P(y|x, s=1) \). The goal is to train \( f_{target} \) while minimizing the best performance of \( f_{inf} \). This is formulated as a minimax optimization problem:

\[
\min_{f_{target}} \left( L_D(f_{target}) + \lambda \max_{f_{inf}} G(f_{inf}) \right)
\]

where \( L_D(f_{target}) \) is the loss function over \( f_{target} \) on \( D \) while \( G(f_{inf}) \) is the gain for \( f_{inf} \) on correctly differentiating between the predictions corresponding to \( s = 0 \) and \( s = 1 \). Empirically, we solve this optimization by following \( k \) epochs of inner maximization to find the best \( f_{inf} \). This is followed by few epochs of solving the outer minimization to find the best \( f_{target} \) with no distinguishable predictions for different values of \( s \). The hyperparameter \( \lambda \) maintains a balance between \( f_{target} \) accuracy and indistinguishable predictions for different values of \( s \). In each iteration, \( \mathcal{A} \)'s strategy is taken into account for enhancing the model performance.

7 EVALUATION

We start by evaluating the effectiveness of prior attribute inference attacks proposed in literature (Section 7.1) before assessing the effectiveness of the proposed attribute inference attacks from the \( \mathcal{A} \)'s perspective (Section 7.2) and \( M \)'s perspective, i.e., DIKAIOS (Section 7.3).

7.1 Baseline: Mahajan et al. and Song et al.

We first evaluate the effectiveness of prior attribute inference attacks proposed in literature closest to our attack methodology by Mahajan et al. [11] and Song et al. [41]. We evaluate the effectiveness of these baseline attacks in both settings: including \( s \) for training ("w/ s") and without including \( s \) for training ("w/o s").

Table 3: Accuracy of baseline attribute inference attacks (Mahajan et al. [11] and Song et al. [41]) show poor performance as they fail to account for the class imbalance in \( s \).

| Dataset  | Recall (Race) | Recall (Sex) | Precision (Race) | Precision (Sex) | F1-Score (Race) | F1-Score (Sex) | Accuracy (Race) | Accuracy (Sex) |
|----------|---------------|--------------|------------------|-----------------|----------------|----------------|----------------|----------------|
| CENSUS   | 0.50          | 0.50         | 0.50             | 0.50            |                |                |                 |                 |
| COMPAS   | 0.62          | 0.50         | 0.70             | 0.52            |                |                |                 |                 |
| MEPS     | 0.55          | 0.55         | 0.56             | 0.56            |                |                |                 |                 |
| LAW      | 0.65          | 0.50         | 0.83             | 0.52            |                |                |                 |                 |
| CREDIT   | 0.50          | 0.53         | 0.50             | 0.65            |                |                |                 |                 |

As seen in Table 3, without including \( s \) for training the model \( f_{target} \) (i.e., corresponding to the more practical threat model), the balanced accuracy of these attacks to correctly infer \( s \) from the model predictions is 0.51 \( \pm 0.02 \) and 0.56 \( \pm 0.06 \) for Race and Sex attributes, respectively. As expected, including \( s \) in the training improves the attack success with 0.61 \( \pm 0.12 \) for Race and 0.55 \( \pm 0.05 \) for Sex. Note that both Sex and Race are binary and hence a random guess corresponds to an accuracy at 0.50. The poor performance of these prior attacks can be attributed to the class imbalance in \( s \) which are not accounted [11, 41]. Indeed, the class imbalance shifts the optimal threshold from the default value of 0.5 used by these attacks to infer the value of \( s \) from \( f_{target}(x) \) or \( f_{target}(x, s) \).

7.2 Evaluation in \( \mathcal{A} \)'s Threat Model (w/o S)

We evaluate our attack in \( \mathcal{A} \)'s threat model where \( \mathcal{A} \)'s only observes \( f_{target}(x) \) to infer \( s \). Table 4 reports results for the two variants of the attack optimizing the AucRoc curves or the PrecRec when \( s \) is not included for training (i.e., "w/o s"). For all datasets, the attack accuracy for both variants is higher than the baseline accuracy for Mahajan et al. [11] and Song et al. [41] (Table 3). Hence, the use of optimal threshold improves the attack performance. In general, the performance of PrecRec variant is higher than the AucRoc variant across all datasets (e.g., an average F1-Score of 0.76 \( \pm 0.06 \) (Race) and 0.76 \( \pm 0.16 \) (Sex) for the PrecRec variant compared to 0.59 \( \pm 0.08 \) (Race) and 0.65 \( \pm 0.14 \) (Sex) for the AucRoc variant).

Table 4: Evaluation of proposed attribute inference attacks from \( \mathcal{A} \)'s perspective where \( s \) is not included in \( D \). The maximum value in each column corresponding to each dataset is indicated in bold.

| Dataset | Recall (Race) | Recall (Sex) | Precision (Race) | Precision (Sex) | F1-Score (Race) | F1-Score (Sex) | AucRoc | PrecRec |
|---------|---------------|--------------|------------------|-----------------|----------------|----------------|--------|---------|
| CENSUS  | 0.50          | 0.50         | 0.50             | 0.50            |                |                | 0.56   | 0.67    |
| COMPAS  | 0.62          | 0.50         | 0.70             | 0.52            |                |                | 0.62   | 0.56    |
| MEPS    | 0.56          | 0.56         | 0.56             | 0.56            |                |                | 0.59   | 0.55    |
| LAW     | 0.65          | 0.50         | 0.83             | 0.52            |                |                | 0.64   | 0.59    |
| CREDIT  | 0.50          | 0.53         | 0.50             | 0.65            |                |                | 0.81   | 0.75    |

7.3 Evaluation in \( M \)'s Threat Model (w/ S)

After showing the attack success of our proposed attack from the \( \mathcal{A} \)'s perspective, we now evaluate from the \( M \)'s perspective where \( s \) is included in the training dataset of the target model. This is a practical assumption since \( M \) has access to \( s \) and aims to estimate the worst case privacy risk by estimating the attack success without any censoring of \( s \). We adapt AucRoc and PrecRec variants in this setting when \( s \) is included in training data. Results of DIKAIOS are reported Table 5.

The attack accuracy for both variants of our attack in \( M \)'s threat model is higher than the baseline accuracy for Mahajan et al. [11]
Table 5: Evaluation of proposed attribute inference attacks from the $M$’s perspective where $s$ is included in training dataset of the target model. The maximum value in each column corresponding to each dataset is indicated in bold.

| Dataset       | Recall | Precision | F1-Score | Accuracy |
|---------------|--------|-----------|----------|----------|
|               | Race   | Sex       | Race     | Sex      | Race     | Sex      |
| CENSUS        | 0.71   | 0.65      | 0.95     | 0.81     | 0.81     | 0.73     | 0.68     | 0.67     |
| COMPAS        | 0.70   | 0.76      | 0.72     | 0.86     | 0.71     | 0.81     | 0.73     | 0.61     |
| MEPS          | 0.40   | 0.36      | 0.53     | 0.64     | 0.45     | 0.46     | 0.60     | 0.56     |
| LAW           | 0.83   | 0.60      | 0.99     | 0.66     | 0.90     | 0.63     | 0.83     | 0.59     |
| CREDIT        | 0.52   | 0.67      | 0.74     | 0.74     | 0.61     | 0.71     | 0.55     | 0.66     |

| Dataset       | Recall | Precision | F1-Score | Accuracy |
|---------------|--------|-----------|----------|----------|
|               | Race   | Sex       | Race     | Sex      | Race     | Sex      |
| CENSUS        | 0.90   | 0.91      | 0.92     | 0.70     | 0.91     | 0.79     | 0.85     | 0.68     |
| COMPAS        | 0.72   | 0.97      | 0.67     | 0.82     | 0.69     | 0.89     | 0.71     | 0.80     |
| MEPS          | 0.57   | 0.91      | 0.50     | 0.54     | 0.53     | 0.68     | 0.64     | 0.54     |
| LAW           | 0.98   | 0.96      | 0.97     | 0.57     | 0.97     | 0.72     | 0.81     | 0.61     |
| CREDIT        | 0.99   | 0.84      | 0.69     | 0.67     | 0.81     | 0.75     | 0.69     | 0.66     |

Table 6: Comparison of prior attribute inference attacks (Fredrikson et al. [14, 15] and Yeom et al. [47]) with Dikaio. These attacks include $s$ in $D$ and compared with Dikaio in same setting. The values in this table, higher than maximum values in Table 5 are indicated in bold.

| Dataset       | Recall | Precision | F1-Score | Accuracy |
|---------------|--------|-----------|----------|----------|
|               | Race   | Sex       | Race     | Sex      | Race     | Sex      |
| CENSUS        | 0.69   | 0.95      | 0.92     | 0.69     | 0.79     | 0.80     | 0.59     | 0.53     |
| COMPAS        | 0.84   | 0.46      | 0.45     | 0.77     | 0.59     | 0.57     | 0.48     | 0.41     |
| MEPS          | 0.66   | 0.74      | 0.73     | 0.78     | 0.69     | 0.76     | 0.76     | 0.75     |
| LAW           | 0.81   | 0.98      | 0.98     | 0.58     | 0.88     | 0.73     | 0.80     | 0.52     |
| CREDIT        | 0.76   | 0.72      | 0.72     | 0.62     | 0.74     | 0.66     | 0.55     | 0.53     |

| Attack (w/) | Recall | Precision | F1-Score | Accuracy |
|-------------|--------|-----------|----------|----------|
| Attack 1: Membership oracle with bounded loss | CENSUS | 0.55 | 0.53 | 0.91 | 0.70 | 0.68 | 0.60 | 0.64 | 0.54 | 0.53 |
|             | COMPAS | 0.47 | 0.53 | 0.45 | 0.83 | 0.46 | 0.64 | 0.49 | 0.51 |
|             | MEPS   | 0.53 | 0.51 | 0.39 | 0.55 | 0.45 | 0.53 | 0.53 | 0.52 |
|             | LAW    | 0.50 | 0.49 | 0.93 | 0.56 | 0.65 | 0.52 | 0.50 | 0.49 |
|             | CREDIT | 0.56 | 0.56 | 0.75 | 0.69 | 0.65 | 0.62 | 0.57 | 0.59 |

| Attack 2: Assuming Normal Distribution | CENSUS | 0.16 | 0.08 | 1.00 | 1.00 | 0.28 | 0.15 | 0.58 | 0.54 |
|                                        | COMPAS | 2e-3 | 0.77 | 0.15 | 0.83 | 4e-3 | 0.79 | 0.50 | 0.51 |
|                                        | MEPS   | 0.07 | 1.00 | 1.00 | 0.59 | 0.15 | 0.74 | 0.54 | 0.60 |
|                                        | LAW    | 0.01 | 0.00 | 0.39 | 0.60 | 0.03 | 0.00 | 0.42 | 0.50 |
|                                        | CREDIT | 0.23 | 0.22 | 0.96 | 0.97 | 0.38 | 0.36 | 0.61 | 0.60 |

| Attack 3: Assuming Uniform Distribution | CENSUS | 0.61 | 0.58 | 0.91 | 0.71 | 0.73 | 0.64 | 0.54 | 0.55 |
|                                        | COMPAS | 0.35 | 0.67 | 0.34 | 0.83 | 0.35 | 0.74 | 0.39 | 0.52 |
|                                        | MEPS   | 0.70 | 0.58 | 0.80 | 0.87 | 0.75 | 0.69 | 0.80 | 0.74 |
|                                        | LAW    | 0.51 | 0.52 | 0.91 | 0.57 | 0.65 | 0.54 | 0.40 | 0.50 |
|                                        | CREDIT | 0.54 | 0.39 | 0.81 | 0.74 | 0.65 | 0.51 | 0.63 | 0.59 |

and Song et al. [41] ("w/ s" in Table 3). In this case too, the use of optimal threshold improves the attack performance. Similar to results of Dikaio from $A$’s perspective (Table 4), we note that AucRoc variant performs better for precision while PrecRec variant results in high recall and F1-score which the variant specifically optimizes for.

Results for Fredrikson et al. [14, 15] and Yeom et al. [47] Attacks. We now evaluate and compare with Fredrikson et al. [14, 15] and Yeom et al. [47] attacks which have the same threat model (as $M$’s for Dikaio) of including $s$ to train the model (w/ $S$). Both these attacks include $s$ in the training dataset for the target model. For a fair comparison, we use Dikaio from the $M$’s perspective where $s$ is included in the training dataset. Table 6 reports results of these prior attribute inference attacks. All the attack success values across different metrics which perform better than Dikaio are indicated in bold. Fredrikson et al. [14, 15] has the best F1-Score of $0.70 \pm 0.08$ (Race) and $0.73 \pm 0.09$ (Sex) compared to all variants of Yeom et al. [47] attacks. However, this is much lower than the PrecRec performance of Dikaio. Out of the three variants of Yeom et al. attacks, the best attack performance was for Attack 3 with F1-Score of $0.62 \pm 0.08$ (Race) and $0.62 \pm 0.14$ (Sex) followed by Attack 1 with F1-Score of $0.58 \pm 0.04$ (Race) and $0.57 \pm 0.10$ (Sex) across all five datasets. We note that Yeom et al.’s Attack 2 has poor F1-Score of $0.40 \pm 0.31$ (Race) and $0.16 \pm 0.14$ (Sex). Clearly, none of the attacks outperform Dikaio consistently across all the metrics (precision, recall, F1-score and accuracy) despite the strong $A$ assumption of knowing the non-sensitive features.

Choice of variant for Auditing. Based on results indicated in Table 7 and 8, the choice of the attack for auditing depends on the application. For critical applications, where accounting for false-negatives is critical, recall is more important metric than precision. In such cases, AucRoc attack for auditing is better than PrecRec attack. On the other hand, for applications where false-negatives are not important, PrecRec attack with a higher recall can be used. For demonstration of Dikaio auditing of algorithmic fairness in in-processing algorithms (Section 8), we use AucRoc attack.

8 PRIVACY AUDITING OF FAIRNESS ALGORITHMS

Having evaluate the effectiveness of our proposed attribute inference attacks to estimate the attribute privacy risks, we use it for conducting an attribute privacy auditing for in-processing group fairness algorithms. Ideally, improving model fairness by adding constraints during training, lowers the model prediction distinguishability and hence the attack accuracy. However, given that class imbalance in $s$ influences the attack success, we hypothesize
Figure 1: Attack accuracy for models with and without in-processing group fairness algorithms (Reductions and AdvDebias) is highly variable on varying the proportion of sensitive attribute (minority class) in dataset.

that varying the proportion of classes for $s$ in $\mathcal{D}$ influences the model's vulnerability to attribute inference attacks. To evaluate
this, we evaluate the attack success on varying the proportion of the minority class in $D$.

On using the fairness algorithms, we indeed find that the disparate impact is above 80% for both Reductions and AdvDebias algorithms (Appendix C), i.e., the models are fair on using the fairness algorithms. In Figure 1, we find the attack success for different proportions of minority class of $s$ is variable. We conjecture that the skewness in the $s$ results in a varying influence on the model predictions resulting in distinguishable output predictions. This in turn impacts the attack accuracy. Furthermore, there is no consistent trend for different datasets which is due to dataset specific properties which warrants more investigation as future work. This additional leakage of sensitive attribute is crucial to be accounted in the model design for model deployment by $M$.

9 RELATED WORK

Privacy Risks via Inference Attacks. ML models leak significant training data information about users. Membership Inference attacks enable an adversary to trace a data point to the model’s training data [7, 20, 25, 26, 34, 38, 40, 42]. A stronger attack is where the adversary can reconstruct the model’s personal and sensitive training data [14, 15]. Property inference attacks have been studied in both blackbox and whitebox setting but with the use of multiple shadow models trained on data with and without property which is used to create synthetic dataset to train an attack model [3, 17]. However, these attacks aim to infer population details of the sensitive attribute instead of the value of the sensitive attribute as in attribute inference attacks. Attribute inference attack is closer to the problem of data amputation where given non-sensitive attributes, the goal is the predict the sensitive attribute as in attribute inference attacks. Attribute inference attack is closer to the problem of data amputation where given non-sensitive attributes, the goal is the predict the sensitive attribute as in attribute inference attacks.

Mitigating Privacy Risks. Several defences have been proposed to mitigate membership inference attacks [1, 12, 23, 33, 39, 43]. For attribute inference, Attriguard [22] leverage adversarial examples and add noise to the output prediction to force the attack model to misclassify to protect sensitive attributes in social networks. Adversarial training to make the predictions over sensitive attributes can be used as a defence [53]. Further, disentangling sensitive features from intermediate input representation through adversarial training [13, 18, 28, 32, 44] and variational auto encoders can be used [49]. These have been shown to be ineffective by Song et al. [41] however within the specific setting where the adversary has access to intermediate features as seen in model partitioning.

Privacy and Fairness. Adversarial training has been explored in context of enhancing fairness quantified [5, 27]. Privacy and fairness have been viewed under the same framework. Training a model with DP results in disproportionate accuracy degradation across minority subgroups [4, 36]. To mitigate the unfairness resulting from differentially private stochastic gradient descent [1], a training algorithm has been proposed to ensure no disparate impact on the groups [45]. Further, a difference in privacy leakage through membership inference attacks were shown to be different for different subgroups in the dataset despite the model showing random guess leakage over the entire data population [46]. Additionally, fair models trained with the constraint of group fairness are more vulnerable to membership inference attacks [6]. The trade-offs and (in)compatibility of fairness, privacy and accuracy have been theoretically studied [9, 37].

10 DISCUSSION AND CONCLUSIONS

In this paper, we propose a novel attribute inference attacks from the model predictions in practical blackbox setting. This attack computes an adaptive threshold to account for imbalance in $s$ which makes it an effective tool for $M$ for privacy auditing fairness algorithms. We adapt the attack for $M$ by including $s$ to train the model which allows $M$ to estimate worst-case attribute privacy risk. We exhaustively evaluated our attack and show that our proposed attribute inference attacks outperforms state-of-the-art approaches and is also effective to evaluate the attribute privacy risk of fairness algorithms.

Note that we assume that $D_{aux}$ used by $A$ or $M$ is from the same distribution as $D$. From the perspective of $M$, this is practical, since $M$ has access to $D$ and using $D_{aux}$ to simulate the adversary behaviour to compute the thresholds. From the $A$’s perspective, if $D_{aux}$ is not from the same distribution as $D$, the threshold computed might not be perfect further impacting the $f_{target}$’s performance. However, this assumption of same data distribution is common across several prior inference attacks [11, 14, 15, 41]. We keep exploration of the effectiveness of our attacks when using $D_{aux}$ from same domain but different distribution as future work.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In ACM SIGSAC Conference on Computer and Communications Security (Vienna, Austria) (CCS ’16). Association for Computing Machinery, New York, NY, USA, 308–318. https://doi.org/10.1145/2976749.2978318
[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80). Jennifer Dy and Andreas Krause (Eds.). PMLR, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html
[3] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. Int. J. Secur. Netw. 10, 3 (Sept. 2015), 137–150.
[4] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In Advances in Neural Information Processing Systems: 15479–15488.
[5] Alex Beutel, Jilin Chen, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-Only Membership Inference Attacks. In Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139). Marina Mehta and Tong Zhang (Eds.). PMLR, 1064–1074. https://proceedings.mlr.press/v139/choquette-choo21a.html
A DISTINGUISHABLE MODEL PREDICTIONS

Figure 2: Distinguishable output predictions for two subgroups of Race and Sex attributes. The adversary exploits this to infer the sensitive attributes for an arbitrary input data record.

We motivate our work by plotting the model predictions across different class values of sensitive attributes. In Figure 2, we observe the output predictions are distinguishable for different values of sensitive attributes (i.e., \( s \in \{0, 1\} \)) even when the attribute was not included in the target model’s training data.
The class imbalance of the sensitive attributes results in the model’s discriminatory behaviour on different classes of the sensitive attribute. For instance, for the CENSUS dataset, the prediction distributions for both RACE and SEX corresponding to the minority group (blacks and female) have a larger peak for lower output probability values. This indicates that the classifier is likely to predict <50K salary for non-white and females compared to members of majority group. For the COMPAS dataset, the model is likely to predict males and blacks to re-offend compared to females or whites. Furthermore, in the MEPS dataset, the utilization of medical resources (visits to medical facility) is predicted as being lower for blacks and females. The model’s discriminatory behaviour results in distinguishable output predictions for the adversary which forms the basis of attribute inference attacks from the model’s prediction [11, 41].

Additionally, we note that there is skew in the model predictions due to class imbalance of sensitive attribute. Hence, the default threshold of 0.5 over the attack model’s predictions will not result in correct inference of sensitive attribute’s value (Figure 2). This motivates our work to design novel attribute inference attacks to account for the skew in the model predictions due to the class imbalance of the sensitive attribute.

## B SANITY CHECK FOR COMPUTING ADAPTIVE THRESHOLD

Before evaluating the effectiveness of the proposed attack variants under Dikaios, we plot the AUCROC and Precision-Recall curves for all the five datasets. In Figure 3, we note that both AUCROC and Precision-Recall curves are beyond the random guess line which indicates the possibility of computing the optimal threshold for attack model trained over imbalanced dataset. LAW dataset only includes RACE attribute since the SEX attribute was close to random guess. This can be explained by the non-distinguishability of model predictions between “Male” and “Female” in Figure 2.

## C FAIRNESS LEVELS FOR THE TWO INPROCESSING ALGORITHMS

| Dataset    | Reductions | Adversarial Debiassing |
|------------|------------|------------------------|
| CENSUS     | 0.95       | 1.00                   |
| COMPAS     | 0.95       | 1.00                   |
| LAW        | 0.98       | 1.00                   |
| MEPS       | 0.97       | 1.00                   |
| CREDIT     | 0.92       | 1.00                   |

Table 7: Disparate Impact for In-Processing Fairness Algorithms for the default proportion on loading the dataset.