Washing The Unwashable : On The (Im)possibility of Fairwashing Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

The use of black-box models (*e.g.*, deep neural networks) in high-stakes decision-making systems, whose internal logic is complex, raises the need for providing explanations about their decisions. Model explanation techniques mitigate this problem by generating an interpretable and high-fidelity surrogate model (*e.g.*, a logistic regressor or decision tree) to explain the logic of black-box models. In this work, we investigate the issue of fairwashing, in which model explanation techniques are manipulated to rationalize decisions taken by an unfair black-box model using deceptive surrogate models. More precisely, we theoretically characterize and analyze fairwashing, proving that this phenomenon is difficult to avoid due to an irreducible factor—the unfairness of the black-box model. Based on the theory developed, we propose a novel technique, called FRAUD-Detect (FaiRness AUDit Detection), to detect fairwashed models by measuring a divergence over subpopulation-wise fidelity measures of the interpretable model. We empirically demonstrate that this divergence is significantly larger in purposefully fairwashed interpretable models than in honest ones. Furthermore, we show that our detector is robust to an informed adversary trying to bypass our detector.

1 Introduction

The wide applicability of machine learning models has recently increased their usage in high-stakes decision systems such as credit scoring [43], insurance risk [10] and predictive justice [30]. The consequences of erroneous decisions that are based on predictions of machine learning models (*e.g.*, people being wrongly denied parole [46]) have increased the demand—from both the public and government—to provide an explanation to humans about model decisions. For instance, this appears as an explanation requirement in the European General Data Protection Regulation [26]. However, various widely-used model architectures, such as deep neural networks, are considered black-boxes due to their complex and hidden internal logic, impeding the ability to explain their decisions in terms that are understandable by a human. To address this issue, black-box model explanation techniques [27, 6] aim to provide reasons for predictions of black-box models in human-understandable terms. One possible way to realize this is to approximate the black-box model using an inherently interpretable model (*e.g.*, logistic regression or decision tree) whose decisions can be easily explained by design [27]. For instance, in logistic regression, the coefficients of the model directly represent the importance of each input feature, while a decision tree is composed of a set of explainable if-then decision rules related to the input features.

In this black-box model explanation setting, the interpretable model is trained with the objective that its predictions agree with the predictions of the black-box model (*i.e.*, high fidelity [20]). However, black-box model explanations can be manipulated to provide deceiving explanations, even if these explanations display a high fidelity with respect to the original black-box model. An example of such ma-

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

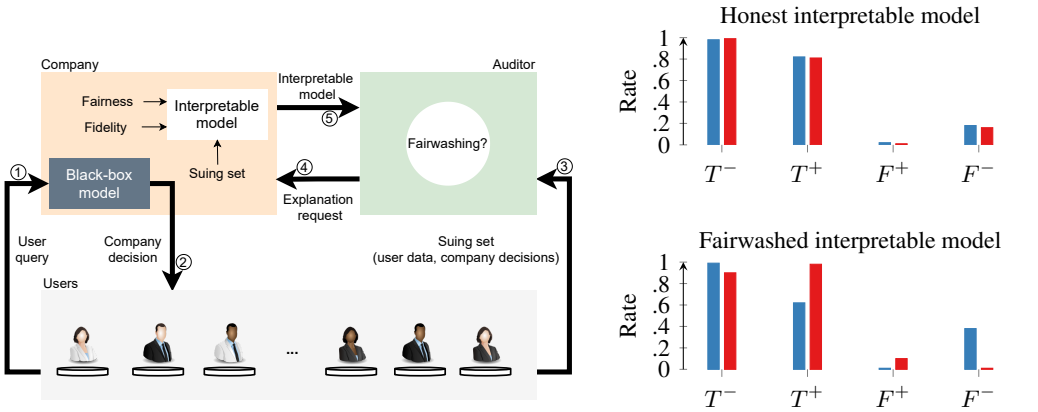

Figure 1: A dishonest company provides a high-fidelity but fairwashed interpretable model to rationalize decisions made by their unfair black-box model. FRAUD-Detect can be exploited by the auditor to distinguish between the fairwashed and honest interpretable model in a non-cooperative way with only query access to models. FRAUD-Detect detects fairwashing based on the agreement and disagreement of the interpretable model with the black-box model among subpopulations 0 ( ) and 1 ( ) quantified by the subpopulation-wise confusion matrix including the True Positive (T^+), False Positive (F^+), False Negative (F^-) and True Negative (T^-) rate of the interpretable model w.r.t. the black-box predictions. The top (bottom) plot shows that the confusion matrix of subpopulations are similar (different) in the honest (fairwashed) interpretable model.

nipulation occurs when the explanations provided cover up the unfairness of the underlying black-box model, resulting in *fairwashing* [2, 3]. As a concrete example, a dishonest bank [4] could use a black-box model that denies a loan to a customer in a discriminative manner—based on a sensitive attribute such as race or gender identity—while providing an interpretable surrogate model promoting the false impression that the decisions are fair and based solely on non-sensitive features. These manipulations are made possible in part by the ambiguity in current regulations [26] when it comes to describing what constitutes a valid explanation [45]. Fairwashing is clearly an ethical issue as individuals who have unfairly received unfavorable decisions are deprived of a chance to contest the decision [3].

Due to the growing importance of fairness in machine learning [7], we focus on a scenario in which a dishonest entity attempts to perform fairwashing [2, 3] when providing model explanations to an auditor (*e.g.*, an external dedicated party). For example, if a company's users expect that the decisions they received were a result of model bias, they may send their data and received decisions to an auditor. The auditor may make a legal demand [8] on the company to provide an interpretable model to explain these users' received decisions (see Figure 1). In this case, a dishonest company has an incentive to manipulate the black-box model explanation technique, to under-report and occlude the unfairness of the black-box model under scrutiny to evade the potential consequences of unfair decision-making. The dishonest entity could thus mislead the auditor by rationalizing its decisions using the corresponding fairwashed interpretable model. In this paper, we are the first to propose a method for detecting such dishonest model explanations. This is a challenging problem. Fairwashing cannot be detected through computing the explanation violation of the explainable model, such as fidelity of explainable models, as we demonstrate that the explainable model cannot achieve perfect fidelity with respect to the black-box model. In addition to this, measuring the differences between the fairness of the black-box model and the explainable model cannot help to detect fairwashing as there are several design choices (including fairwashing) that could lead to different fairness, thus non-intentional fairwashing.

Our primary contributions include (1) an extensive theoretical analysis of fairwashing and the fairness limitations in explanation techniques and (2) a novel approach for detecting, and thus deterring, fairwashing. Our proposed method, FRAUD-Detect (FaiRness AUDit Detection), formulates the problem of detecting fairwashing as a non-cooperative test; both to overcome possible dishonest behaviors in entities that are actively being audited and for communication efficiency. FRAUD-Detect operates in a realistic scenario: FRAUD-Detect does not require access to the black-box model provided the users calling for the audit have provided their data and received decisions, which is a realistic setting as often the model is not provided by the entity due to intellectual property and trade-secret concerns [39, 37]. Thus, FRAUD-Detect only relies on the predictions of the interpretable

and black-box models. Both our theoretical and empirical analysis quantify the per-subpopulation, per-label fidelity of the interpretable model with respect to the black-box model.

Figure 1 shows confusion matrix distributions for two sensitive subpopulations before and after fairwashing. These matrices are constructed by comparing the interpretable to the black-box model predictions. Fairwashing induces a divergence between subpopulation values to conceal the unfairness of the black-box model. FRAUD-Detect leverages the Kullback–Leibler (KL) divergence between subpopulation confusion matrices to distinguish between honest and fairwashed interpretable models. For a comprehensive empirical study, we examine logistic regressors and decision trees as interpretable models with respect to black-box Deep Neural Networks, AdaBoost [24], Gradient Boosted Decision Trees [17] and Random Forests [13] on three benchmark datasets: COMPAS [5], Adult Income [22] and Bank Marketing [34].

We evaluate the strength of our detector in the presence of an *informed adversary* representing an informed dishonest entity with knowledge of FRAUD-Detect that seeks to fairwash *and* bypass detection by constraining the divergence over subpopulation-wise confusion matrices. Our empirical experiments quantify the capacity of this informed adversary by computing the achievable fairness gap and fidelity under the additional detection constraint. We provide additional theoretical support for the challenge of jointly satisfying fairwashing and evasion in appendices.

In summary, our main contributions are as follows:

- We characterize fairwashing via a theoretical analysis on the difference in subpopulation gap between the black-box model and the interpretable model, *fairwashing violation* (Section 3).
- Based on our fairwashing theory, we establish requirements for a *sufficient* fairwashing detection method and prove that fairwashing is impossible to avoid completely due to an *irreducible* component in the violation term.
- We introduce the first method for fairwashing detection (Section 4.2). Informed by our theoretical results, we observe that fairwashing causes disparate effect on the subpopulation-wise fidelity distributions of the interpretable model with respect to the black-box predictions. We leverage this observation to propose a non-cooperative black-box access method, FRAUD-Detect, which utilizes KL divergence on per-subpopulation confusion matrices to distinguish between fairwashed and honest interpretable models.
- Our empirical results demonstrate that FRAUD-Detect successfully detects fairwashing based on the KL divergence between subpopulation-wise confusion matrices (Section 6). We show that the divergence over subpopulation confusion matrices can vary by over 0.6 between an honest and a fairwashed interpretable models.
- We illustrate the robustness of FRAUD-Detect against an informed adversary (*i.e.*, a dishonest entity who attempts to fairwash while evading detection). Our empirical results show that evading our detector comes at the cost of a significant increase in subpopulation gap, negating fairwashing. Specifically, a dishonest entity that jointly attempts to fairwash while evading detection only achieves parity gaps greater than 10%.

2 Preliminaries and Problem Formulation

2.1 Related Work

Fairness. Many different formal definitions of algorithmic fairness have been proposed in the machine learning community [35]. One of the main challenges of algorithmic fairness is the lack of consensus on a universally applicable fairness definition. More precisely, it has been formally demonstrated that many such measures are incompatible with each other (*i.e.*, they cannot be achieved jointly in some situations) [31]. In addition, their differences are also rooted in their underlying philosophical and moral assumptions [28]. Thus, the choice of a particular fairness metric is usually context dependent. Nonetheless broadly speaking, there are two main families of fairness definitions: individual fairness and group fairness [23]. Individual fairness posits that "similar individuals should be treated similarly". In contrast, group fairness strives to equalize statistical properties of classification outcomes across subpopulations created by partitioning the population based on a sensitive attribute A such as gender. In our work, we focus on a particular notion of group fairness, demographic parity, also called statistical group parity, which refers to observing equal probability of positive label prediction over

subpopulations. One of the reasons we rely on demographic parity as the fairness metric is that true labels will not be available in a realistic scenario (especially in fairwashing scenario described in Section 2.3). In general, the existence of true labels conflicts with the assumption that when requesting decisions (or explanations), users (or auditors) do not have access to these true labels [42]. Furthermore, using the predictions of a black-box model as the true labels is paradoxical in the sense that the black-box model is assumed to be unfair. Demographic parity enforces independence between the class predicted by the model and inclusion in a particular subpopulation.

Definition 1 (Demographic parity [15])**.** *A classifier $\hat{Y}$ satisfies demographic parity with respect to sensitive attribute A if*

$$\Pr[\hat{Y} = 1 | A = a] = \Pr[\hat{Y} = 1 | A = b] \qquad \forall a, b \in A.$$

Global black-box explanation methods focus on explaining the whole logic of black-box models by training an inherently interpretable surrogate model. We refer the interested reader to Appendix A for the description of other explanation methods. In terms of abusing explanations via fairwashing, Fukuchi et al., similarly to us, try to detect fairwashing but with a different setting [25]. Slack et al. shows an alternate method of fairwashing that evades local model explanation techniques in a setting that is comparable to ours but they do not attempt to detect fairwashing [44]. We explore these additional fairwashing settings in Appendix B.

2.2 Desiderata for Global Black-box Explanations

We consider the setting in which an entity learns a complex black-box model $B(\cdot)$ on a training set $(X_{\text{Tr}}, Y_{\text{Tr}}, A)$, in which A represents a sensitive attribute. We maintain the setup from recent fairwashing literature [3, 2] and consider binary classifiers mapping M features into a binary label $B : \mathbb{R}^M \mapsto \{0, 1\}$. We also consider binary-valued sensitive attributes, $A \in \{0, 1\}$. Note that we do not assume that A is necessarily used during the training of $B(\cdot)$. Given user queries, the entity uses the predictions of $B(\cdot)$ as part of a high-stakes decision system. Later, a group of these users may request an explicit explanation from the entity due to concerns over improper use of the sensitive attribute. Such a scenario frequently arises in high-impact domains, such as bank loans and credit scoring. Let X_{sg} be a suing set comprising the unlabeled data of users demanding an explanation for their particular outcomes (see Figure 1). Consequently, an external auditing entity requires the company to provide explanations in terms that are understandable to humans about the predictions of the black-box. Here, we will assume that the explanation will be provided in the form of a global explanation. This means that a simple interpretable model $I(\cdot)$ will be trained on a dataset X labeled by querying the black-box model $B(\cdot)$ trained on X_{Tr}, such that $I(\cdot)$ accurately reflects and explains the logic of $B(\cdot)$ in terms that are understandable to humans.

Formalizing further the auditing desiderata for black-box model explanations, the interpretable model $I(\cdot)$ must accurately mirror: 1) the output predictions of the black-box model (*i.e.*, fidelity) and 2) the fairness (according to a pre-defined metric) of the black-box model.

Fidelity. The interpretable model should satisfy the fidelity criterion for any $X \sim \mathfrak{D}$. Fidelity (defined in Appendix C) is difficult to perfectly achieve and challenging to measure for an auditor over all $X \sim \mathfrak{D}$. Thus, we additionally introduce the notion of empirical fidelity.

Definition 2 (Empirical Fidelity)**.** *The empirical fidelity on a dataset $X = \{x_i\}_{i=1}^N$ including N data points x_i is defined as the relative accuracy of $I(\cdot)$ with respect to $B(\cdot)$ on X:*

$$\texttt{EmpiricalFidelity}(I, B; X) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(I(x_i) = B(x_i)), \tag{1}$$

where the indicator $\mathbb{1}(\cdot, \cdot)$ outputs 1 if the output label of $I(\cdot)$ and $B(\cdot)$ are the same and 0 otherwise.

Fairness. As the audit of $B(\cdot)$ is expressly requested due to fairness concerns with respect to A, the interpretable model should reflect the fairness violations or adherences of the black-box model.

2.3 Fairwashing Definitions

To evade legal consequences of decision-making based on improper use of a sensitive attribute, a dishonest company may perform *fairwashing*. Fairwashing with respect to a pre-defined fairness

metric permits the dishonest company to learn an interpretable model that hides the unfair behaviour of their black-box model. In practice, fairwashing could also occur due to the fact that current regulations [26] do not define what constitutes a *valid explanation*, thus leaving the possibility for the model provider to choose the interpretable model that meets their needs [45]. Fairwashing can be quantified with respect to a particular fairness measure.

Definition 3 (Fairwashing)**.** *Let Γ_I and Γ_B define the fairness gaps of the interpretable model and black-box model, respectively, with respect to a pre-defined fairness metric. For example, in terms of demographic parity (Definition 1):*

$$\Gamma_I := \Pr\left[\hat{Y}_I = 1 \mid A = 0\right] - \Pr\left[\hat{Y}_I = 1 \mid A = 1\right], \Gamma_B := \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] - \Pr\left[\hat{Y}_B = 1 \mid A = 1\right]. \tag{2}$$

We define the fairwashing violation $\gamma > 0$ *as the difference between Γ_I and Γ_B:*

$$\gamma := \Gamma_B - \Gamma_I. \tag{3}$$

Assumption 1. *The fairwashing violation γ is non-negative. Remark that the case of $\gamma < 0$ is the opposite of fairwashing since the interpretable model is displaying a bigger fairness gap than that of the black-box model, thus a model owner has no incentive to employ such an interpretable model.*

3 Fairness Gap of the Black-Box Breeds Fairwashing

We analyze whether it is possible to eliminate the risk of fairwashing altogether. We characterize the fairwashing violation γ and show that completely eliminating fairwashing is impossible: an interpretable model explaining an unfair black-box model always has a non-zero fairwashing violation.

Theorem 1. *Assume $\hat{Y}_B$ are the black-box model $B(\cdot)$ predictions, $\hat{Y}_I$ are the predictions of an interpretable model $I(\cdot)$ explaining $B(\cdot)$ and A is a sensitive attribute: 1) If $B(\cdot)$ does not satisfy demographic parity, completely eliminating fairwashing is impossible (i.e., $\gamma > 0$); 2) A detector for fairwashing measuring false-positive rates and true-positive rates of $I(\cdot)$ with respect to $B(\cdot)$ is sufficient.*

We provide a proof sketch here (see Appendix D for full proof).

Proof Sketch. Define T_a^+, the true-positive rate of $I(\cdot)$ with respect to $B(\cdot)$ on $X \sim \mathfrak{D}_a$, in which $\mathfrak{D}_a$ denotes the distribution over data with attribute $a \in A$. Define F_a^+, the false-positive rate of $I(\cdot)$ with respect to $B(\cdot)$ on $X \sim \mathfrak{D}_a$. Denote the differences between T_0^+ and T_1^+, and F_0^+ and F_1^+ as $\tilde{\delta}, \delta'$, respectively. We can express Γ_I in terms of $\tilde{\delta}, \delta'$, T_1^+, F_1^+, Γ_B and $Y_B|A$ by expanding terms and using Bayes' formula:

$$\Gamma_I = \Gamma_B \left(T_1^+ - F_1^+\right) + \left(\tilde{\delta} - \delta'\right) \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] + \delta' \tag{4}$$

Now that we have derived Γ_I in terms of our desired terms, we can eliminate Γ_I from the equation for γ. The resulting formula for γ in terms of $\tilde{\delta}, \delta'$, T_1^+, F_1^+, Γ_B and $Y_B|A$ allows us to demonstrate **correctness**, **irreducibility** and **sufficiency**.

$$\gamma = \Gamma_B(1 + F_1^+ - T_1^+) - \delta' \Pr\left[\hat{Y}_B = 0 \mid A = 0\right] - \tilde{\delta} \Pr\left[\hat{Y}_B = 1 \mid A = 0\right]. \tag{5}$$

Correctness. Note that if $\Gamma_B = 0$ (*i.e.*, black-box model is fair) then no-fairwashing is present, and no detector should mistakenly report fairwashing. We can verify this with Equation (13). Under Assumption 1, we know that $\gamma \geq 0$ and as below terms are, by definition non-negative, thus:

$$\gamma = -\left(\delta' \Pr\left[\hat{Y}_B = 0 \mid A = 0\right] + \tilde{\delta} \Pr\left[\hat{Y}_B = 1 \mid A = 0\right]\right) \geq 0 \Longrightarrow \tilde{\delta} = \delta' = 0.$$

Irreducibility. If $\Gamma_B \neq 0$, even if $\tilde{\delta} = \delta' = 0$, *i.e.* when true-positive and false-positive rates are equal over subpopulations ($T_0^+ = T_1^+$ and $F_0^+ = F_1^+$), there exists an irreducible fairwashing violation:

$$\gamma = \Gamma_B(1 + F_1^+ - T_1^+) > 0, \tag{6}$$

since $1 + F_1^+ - T_1^+ > 0$ except in trivial cases (see Appendix D for the proof). We further point out that $1 + F_1^+ - T_1^+$ is not a function of disparity in the interpretable model as the value does

not measure disparity between subpopulations. Rather, $1 + F_1^+ - T_1^+$ is a function of the difference in true-positive and false-positive rates (T_1^+ and F_1^+, respectively) within a single subpopulation. The sole measure of disparity in the gap above is black-box Γ_B, which is a given constant.

Sufficiency. The remaining factors of the gap are only function of $\tilde{\delta}$ and δ' as $\Pr\left[\hat{Y}_B = 1 \mid A = 0\right]$ and $\Pr\left[\hat{Y}_B = 1 \mid A = 0\right]$ are also given constants. Therefore, a detector that measures $\tilde{\delta}$ and δ' is sufficient to detect fairwashing. □

4 Detecting Fairwashing

4.1 Fairwashing In Practice

Eliminating the fairness gap of the black-box model would be the ideal outcome as it would also remove the possibility for a dishonest model provider to perform fairwashing. However, in practice absolute fairness is often unrealistic and even regulatory bodies tolerate small violations of fairness laws [18]. In Section 3, we showed that fairness violations result in fairwashing violations unless the interpretable model has a higher fairness gap than the black-box model, which would not be desirable from the point of view of the model provider. Therefore, not all fairwashing violations may have been intended by the model owners. These *involuntary* violations are in fact side-effects of the imperfect fidelity, which is inherent to optimizing the empirical fidelity (Definition 2) of the interpretable model on a training set $X_{\text{tr}} \sim \mathfrak{D}$, $\max \quad \texttt{EmpiricalFidelity}(I, B; X_{\text{tr}})$.

In contrast, deliberate fairwashing is the result of a similar optimization problem purposefully constrained to provide better fairness on a target suing set $X_{\text{sg}} \sim \mathfrak{D}$:

Definition 4 (Fairwashing Optimization [2])**.** *Given a black-box model $B(\cdot)$ and a suing set X_{sg}, fairwashing optimization is defined as learning an interpretable model $I(\cdot)$ from $B(\cdot)$ on X_{sg} such that the interpretable model has 1) high fidelity with respect to the black-box model and 2) is less unfair than this black-box model:*

$$\begin{aligned} \max \quad & \texttt{EmpiricalFidelity}(I, B; X_{sg}) \\ \textit{subject to} \quad & \texttt{FairnessGap}(I; X_{sg}) \leq \epsilon, \end{aligned} \tag{7}$$

in which ϵ is an upper bound on the fairness gap of the interpretable model on the suing set $\texttt{FairnessGap}(I; X_{sg})$, and $\texttt{FairnessGap}(\cdot, \cdot)$ constitutes a measure of fairness gap according to some fairness metric – generically measured by the subpopulation gap in a fairness metric. Note that $\epsilon < \texttt{FairnessGap}(B; X_{sg})$ must be less than the fairness gap of the black-box model on the suing set $\texttt{FairnessGap}(B; X_{sg})$ to cause fairwashing violation $\gamma > 0$ (see Definition 3).

4.2 Proposed method: FRAUD-Detect

As demonstrated in Section 3, a fairwashing detection method that relies on the difference over subpopulations in true-positive and false-positive rates of the interpretable model w.r.t the black-box model ($\tilde{\delta}$ and δ' respectively) is sufficient. Inspired by this theoretical analysis, we propose our fairwashing detector, FRAUD-Detect, which locates fairwashing violations. FRAUD-Detect aims to distinguish between honest and fairwashed interpretable models approximating the unfair black-box model $B(\cdot)$ as described in Section 2.2 under the additional following constraints: 1) without white-box access to the black-box and interpretable models as well as; 2) independently of the cooperation of the model provider being audited. These constraints are introduced to conform to realistic scenarios in which entities desire to retain the confidentiality of their black-box model due to intellectual property rights and because they considered it as a company's asset.

Similarly to definitions of true-positive T_a^+ and false-positive F_a^+ rates of $I(\cdot)$ w.r.t $B(\cdot)$ on data with attribute $a \in A = \{0, 1\}$, we define the true-negative rate, T_a^-, and false-negative rate, F_a^- as: $T_a^- = \Pr[\hat{Y}_I = 0 \mid \hat{Y}_B = 0, A = a]$, $F_a^- = \Pr[\hat{Y}_I = 0 \mid \hat{Y}_B = 1, A = a]$ for $a \in A = \{0, 1\}$. Note that $T_a^- = 1 - F_a^+$ and $F_a^- = 1 - T_a^+$.

FRAUD-Detect computes the $T_a^+, F_a^+, T_a^-, F_a^-$ for each subpopulation $a \in A = \{0, 1\}$ using the predictions of the interpretable model and black-box model on the suing set X_{sg}. Let $C_0 = [T_0^+, F_0^+, T_0^-, F_0^-]$ and $C_1 = [T_1^+, F_1^+, T_1^-, F_1^-]$ be the (flattened) confusion matrices of the

Algorithm 1 FRAUD-Detect. $\{\texttt{predicate}\} = \{(x, a, \hat{y}_B, \hat{y}_I) \in (X_{\text{sg}}, A, \hat{Y}_B, \hat{Y}_I) : \texttt{predicate}\}$

Input: Query access to the interpretable model $I(\cdot)$, suing dataset X_{sg}, Black-box model predictions on the suing set $B(X_{\text{sg}})$, sensitive attribute $a \in A$ and a threshold $\Delta > 0$.

Output: $\begin{cases} T & \text{fairwashing is detected} \\ F & \text{fairwashing is not detected} \end{cases}$

$\hat{Y}_I \leftarrow I(X_{\text{sg}})$ ▷ Query interpretable model
for $i \leftarrow 0, 1$ **do**
  $T_i^+ \leftarrow \frac{|\{\hat{y}_B=1, \hat{y}_I=1, a=i\}|}{|\{\hat{y}_B=1, a=i\}|}$ ▷ TPR
  $F_i^+ \leftarrow \frac{|\{\hat{y}_B=0, \hat{y}_I=1, a=i\}|}{|\{\hat{y}_B=0, a=i\}|}$ ▷ FPR
  $F_i^- \leftarrow \frac{|\{\hat{y}_B=1, \hat{y}_I=0, a=i\}|}{|\{\hat{y}_B=1, a=i\}|}$ ▷ FNR
  $T_i^- \leftarrow \frac{|\{\hat{y}_B=0, \hat{y}_I=0, a=i\}|}{|\{\hat{y}_B=0, a=i\}|}$ ▷ TNR
$C_0 \leftarrow [T_0^+, F_0^+, T_0^-, F_0^-]$
$C_1 \leftarrow [T_1^+, F_1^+, T_1^-, F_1^-]$
$\mathcal{C}_{KL} \leftarrow \text{KL}(C_0, C_1)$ ▷ Kullback–Leibler
if $\mathcal{C}_{KL} > \Delta$ **then**
  return T
else
  return F

interpretable model w.r.t the black-box model on X_{sg} with attribute 0 and 1, respectively. Then, FRAUD-Detect computes the divergence between C_0 and C_1 using Kullback–Leibler (KL) as:

$$\mathcal{C}_{\text{KL}} = \text{KL}(C_0, C_1). \tag{8}$$

Recall that we demonstrated in Section 3 that the divergence in T_0^+ and T_1^+ ($\tilde{\delta}$) and F_0^+ and F_1^+ (δ') were sufficient for a detection method. We can complement[1] our detection method with additional divergences in T_a^- and F_a^- and measure the dissimilarity via KL divergence. This choice is natural as the KL divergence is commonly used to quantify the divergence over probability distributions—C_a functions as a simplified representation of the probability distribution $Y_I|Y_B, A = a$. For an honest interpretable model, $\mathcal{C}_{\text{KL}}$ is generally relatively low ($\tilde{\delta}, \delta' \approx 0$) as the only divergence arises from general error in fidelity optimization and the fairness gap of the black-box model (as shown in Section 3). In other words, the honest interpretable model approximates the black-box model equivalently across subpopulations. However, for a fairwashed interpretable model, $\mathcal{C}_{\text{KL}}$ grows significantly, as the interpretable model is explicitly manipulated to optimize the fairness across subpopulations and improve over the black-box model. We quantify the distinction between honest and fairwashed interpretable models via a threshold $\Delta > 0$ on $\mathcal{C}_{\text{KL}}$ such that if $\mathcal{C}_{\text{KL}} > \Delta$, the interpretable model is considered fairwashed. Note here that due to the irreducibility result stated in Section 3, Δ must be chosen such that Δ is tightly greater than the irreducible term in order to properly distinguish between accidental fairwashing that occurs as a result of the irreducibility and malicious fairwashing. This procedure is detailed in Algorithm 1.

5 Evading FRAUD-Detect

In this section, we investigate on whether a dishonest entity could evade FRAUD-Detect while performing fairwashing. We assume that this dishonest entity, which we call the *adversary*, is informed about FRAUD-Detect. Namely, the adversary is aware of the fairwashing detection method and desires to fairwash while evading detection by the auditor. To achieve this goal, the informed adversary has the objective of finding a fairwashed interpretable model that satisfies an additional constraint on the Kullback–Leibler (KL) divergence of the confusion matrices among subpopulations, $\mathcal{C}_{\text{KL}}$ in the previous section. To probe the robustness of our detector to an informed adversary empirically, we explore the range of fairness gap given a fixed value of fidelity and a fixed value of $\mathcal{C}_{\text{KL}}$ via solving the informed adversary optimization problem.

Definition 5 (Informed Adversary Optimization)**.** *Given a black-box model $B(\cdot)$, a suing set X_{sg}, a sensitive attribute A, a loss threshold v and a fairwashing detection threshold Δ, an informed adversary aims to learn an interpretable $I(\cdot)$ such that*

$$\begin{aligned} &\textit{minimize} && \texttt{FairnessGap}(I; X_{sg}) \\ &\textit{subject to} && L(I; X_{sg}) \leq v \quad \textit{and} \quad \mathcal{C}_{KL} \leq \Delta, \end{aligned} \tag{9}$$

where $\mathcal{C}_{KL}$ is the KL divergence between the subpopulation confusion matrices of X_{sg}.

[1] We performed experiments in two settings: 1) considering all related true-positive, false-positive, true-negative and false-negative rates; 2) considering only independent true-positive and false-positive rates.

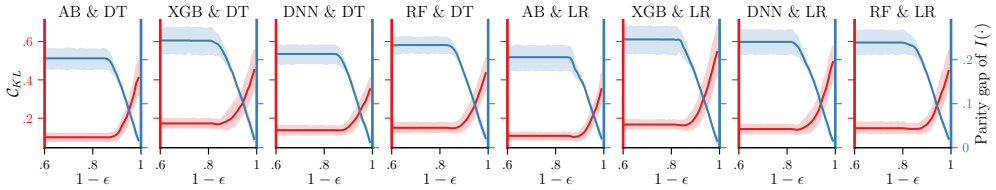

Figure 2: KL divergence between the confusion matrices of subpopulations, $\mathcal{C}_{\text{KL}}$, and parity gap as a function of ϵ fairwashing using COMPAS. Values before $1-\epsilon = .6$ are constant and not shown. See Appendix F for the results of other two datasets.

The differences between the fairwashing optimization in Definition 4, and the informed adversary optimization of Definition 5 are two fold. First, the order of the maximand and the constraint is reversed. Both formulations are valid since each indicates priorities of the adversary, namely, whether to put a hard limit on the fidelity loss or on the fairness gap of the interpretable model. Second, in Definition 5 we replaced the non-differentiable $\texttt{EmpiricalFidelity}(I, B; X_{\text{tr}})$ with a differentiable measure of loss $L(I; X_{sg})$. In our empirical evaluations, we use the logistic regression loss.

To solve the above optimisation problem, we consider the Rashomon set of high-fidelity interpretable models and compute the range of the fairness gap of interpretable models. In short, given a classification task, the Rashomon Set [41] is defined as the set of almost-equally-accurate models (interpretable models in our case). More precisely, given a model class $\mathcal{F}$, a loss function $L_{\mathcal{D}}(\cdot)$ over a dataset $\mathcal{D}$ of interest, a reference model f^* (*e.g.*, optimal model) and a performance threshold $\tau \in [0, 1]$, the Rashomon set $R_s(\mathcal{F}, f^*, \tau) = \{f \in \mathcal{F} \mid L_{\mathcal{D}}(f) \leq L_{\mathcal{D}}(f^*) + \tau\}$.

We extend the Fairness in the Rashomon Set (FaiRS) algorithm [19] to compute the range of the fairness gap of interpretable models that can be generated over the set of high-fidelity models satisfying a constraint on the KL divergence. FaiRS exploits the so-called Rashomon Effect [14], which is an empirical phenomenon resulting in multiple models displaying the same performance overall (*e.g.*, w.r.t their global accuracy) but have significant differences in terms of their individual predictions [19]. In our setting, the Rashomon effect implies that several interpretable models can achieve the same high fidelity w.r.t the black-box model while displaying different values of fairness gap. Thus, computing the range of the fairness gap of high-fidelity interpretable models that satisfy a constraint on $\mathcal{C}_{\text{KL}}$ quantifies the robustness of FRAUD-Detect to the informed adversary.

To the best of our knowledge, the FaiRS algorithm of [19] is the only work proposing a practical solution to efficiently characterize the range of the fairness gap over the Rashomon set. Alternative approaches would require measuring the range of the fairness gap over approximations of the Rashomon set obtained by either generating models using several hyperparameter values (*e.g.*, seeds [33, 21]) or brute-forcing over a particular class of model (*e.g.*, depth seven decision trees [41]). This is not practical due to the high computational cost and may lead to sub-optimal results.

6 Empirical validation

FRAUD-Detect aims to detect fairwashing by distinguishing between a fairwashed interpretable model and honest interpretable one during an audit. Additionally, FRAUD-Detect seeks to be robust even against an informed adversary, *i.e.*, demonstrate the ability to detect fairwashing even when a dishonest entity is aware of FRAUD-Detect and attempts evasion. Therefore, we empirically validate the performance of FRAUD-Detect in: 1) capturing a relationship between $\mathcal{C}_{\text{KL}}$ and demographic parity gap of the interpretable model $I(\cdot)$: $\mathcal{C}_{\text{KL}}$ increases as the demographic parity gap of the interpretable model decreases (*i.e.* as fairwashing becomes more severe); 2) restricting the fairwashing capabilities of the informed adversary: incorporation of $\mathcal{C}_{\text{KL}}$ constraint into fairwashing problem prevents favorable demographic parity gaps for the interpretable model, precluding fairwashing.

We assess the performance of FRAUD-Detect using a diverse set of black-box architectures, interpretable models and datasets. More precisely, we consider four architectures of black-box models: Deep Neural Networks (DNN), AdaBoost (AB) [24], Gradient Boosted Decision Trees (XGBs) [17] and Random Forests (RFs) [13]. We evaluate the approach on three real-world datasets corresponding to critical decision systems: Adult Income [22], Bank Marketing [34] and COMPAS [5]. We rely on Logistic Regression (LR) and Decision Trees (DTs) as our interpretable models due to their high performance and ease of explainability. We refer to Appendix E for details on these datasets, black-box models, interpretable models and implementation of FRAUD-Detect.

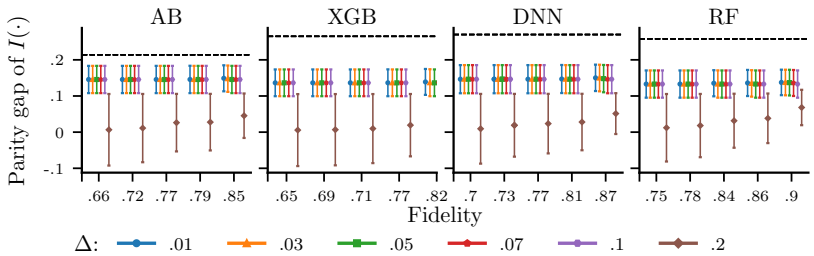

Figure 3: Range of the demographic parity gap of fairwashed logistic regression interpretable models subjected to a constraint (Δ) on the $\mathcal{C}_{\text{KL}}$ using COMPAS. Horizontal lines denote the parity gap of the black-box models. See Appendix F for the results of other two datasets.

FRAUD-Detect successfully detects fairwashing. Figure 2 illustrates both the demographic parity gap and $\mathcal{C}_{\text{KL}}$ of interpretable models as a function of ϵ. The bigger the value of ϵ, the stronger the fairwashing (see Definition 4). Any graph within the figures may be read from left to right as the degree of fairwashing increases. As fairwashing becomes more severe, the demographic parity gap decreases while $\mathcal{C}_{\text{KL}}$ increases such that we observe a relationship between the demographic parity gap of interpretable models and the associated $\mathcal{C}_{\text{KL}}$ values. For example, decreasing the demographic parity gap of Decision Trees trained on a COMPAS dataset labelled by a RF black-box model by half, increases $\mathcal{C}_{\text{KL}}$ by a factor of 4. The amount of change in $\mathcal{C}_{\text{KL}}$ and in the demographic parity gap of fairwashed interpretable models differ across the dataset and black-box models due to the different original fairness of the black-box models. As $\mathcal{C}_{\text{KL}}$ is sensitive to changes in fairwashing, FRAUD-Detect can be used to detect fairwashing by not permitting $\mathcal{C}_{\text{KL}}$ to pass a certain threshold value. The $\mathcal{C}_{\text{KL}}$ thresholds corresponding to 5% and 50% fairwashing are reported in Appendix G. Per-seed results are provided in Appendix H.

FRAUD-Detect is robust to the informed adversary. Figure 3 shows the range of the demographic parity gap of high-fidelity fairwashed interpretable models subjected to a constraint on the $\mathcal{C}_{\text{KL}}$ by solving the optimization problem of Equation (9). We use four types of black-box models (AB, XGB, DNN and RF). However, for simplicity, we performed the experiment using logistic regression as the interpretable model. The fairwashing detection thresholds were chosen from a wide range, *i.e.* $\Delta \in \{0.01, 0.03, 0.05, 0.07, 0.10, 0.20\}$ (the constraint on $\mathcal{C}_{\text{KL}}$ in Equation (9)). A direct implication for FRAUD-Detect is that using low fairwashing detection thresholds makes fairwashing difficult. Consistently over all these results, imposing a constraint on the $\mathcal{C}_{\text{KL}}$ significantly narrows the range of demographic parity gap of fairwashed interpretable models. For instance, on COMPAS with $\Delta = 0.2$, for an AB model that has demographic parity gap of 0.214, fairwashing can produce a completely fair (*i.e.*, demographic parity gap of 0.0) high-fidelity interpretable model while remaining undetected (*i.e.*, $\mathcal{C}_{\text{KL}} < 0.2$). However, when $\mathcal{C}_{\text{KL}}$ decreases to 0.1, the fairest high-fidelity interpretable model obtained by the adversary exhibits a demographic parity gap of 0.107.

7 Conclusion and Future Work

Incorporating fairness and explainability in decision systems can help foster trust [29] in them by characterizing failure modes (*e.g.*, unfairness) and providing assurances (*e.g.*, fairness constraints) [16]. Auditing provides an oversight component and helps avoid *first-order failures* such as unfairness. In this sense, fairwashing can be seen as failure of auditing—a *second-order failure* that reduces trust in both fairness and explainability. Second-order failures help characterize the limits of techniques designed to reduce the risks of first-order failures. Our work is the first to delineate the theoretical limits of what is possible in auditing fairness and in containing the risk of fairwashing. In practice, we demonstrated the possibility of performing fairwashing through joint optimization of fidelity and fairness constraints when providing explanations. This could thwart auditing processes that are solely based on an analysis of the delivered interpretable model. We address this issue by proposing an additional auditing protocol that queries the interpretable model. Finally, through both a theoretical framework and an experimental evaluation, we demonstrate that our fairwashing detector cannot be evaded by an attacker, informed of our detector, without significantly degrading the fidelity of model explanations. Future directions include the extension of the detector to other fairness measures (*e.g.*, measures incorporating ground-truth information) as well as extending the analysis to multi-class setups. However, these would both require additional assumptions on the data manifold (to account for class correlations) to establish similar results as in Section 3.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. Stockholm, Sweden, (July 2018).

[2] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Long Beach, California, USA, (June 2019).

[3] Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. 2021. Characterizing the risk of fairwashing. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*. Sydney, Australia, (December 2021).

[4] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing explanations with off-manifold detergent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Online, (July 2020).

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica*, (May 2016). Retrieved January 15, 2022.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

[7] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104, 671.

[8] Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frènay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29, 2, 149–169.

[9] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report MSR-TR-2020-32. Microsoft, (May 2020). `https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/`.

[10] Noorhannah Boodhun and Manoj Jayabalan. 2018. Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*, 4, 2, 145–154.

[11] Olcay Boz. 2002. Extracting decision trees from trained neural networks. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*. Edmonton, Canada, (July 2002).

[12] Leo Breiman. 2017. *Classification and regression trees*. Routledge.

[13] Leo Breiman. 2001. Random forests. *Machine Learning*, 45, 1, 5–32.

[14] Leo Breiman. 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16, 3, 199–231.

[15] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops (ICDM)*. Miami, Florida, USA, (December 2009).

[16] Varun Chandrasekaran, Hengrui Jia, Anvith Thudi, Adelin Travers, Mohammad Yaghini, and Nicolas Papernot. 2021. Sok: machine learning governance. *arXiv preprint arXiv:2109.10870*.

[17] Tianqi Chen and Carlos Guestrin. 2016. XgBoost: a scalable tree boosting system. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*. San Francisco, California, USA, (August 2016).

[18] U.S. Equal Employment Opportunity Commission. 1979. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. (March 1979). Retrieved 01/15/2022 from `https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines`.

[19] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing fairness over the set of good models under selective labels. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Online, (July 2021).

[20] Mark Craven and Jude Shavlik. 1995. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems (NIPS)*, 8, 24–30.

[21] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

[22] Dheeru Dua and Casey Graff. 2017. UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

[23] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*. Cambridge, Massachusetts, USA, (August 2012).

[24] Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 1, 119–139. ISSN: 0022-0000. DOI: `https://doi.org/10.1006/jcss.1997.1504`. `https://www.sciencedirect.com/science/article/pii/S002200009791504X`.

[25] Kazuto Fukuchi, Satoshi Hara, and Takanori Maehara. 2020. Faking fairness via stealthily biased sampling. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)* number 01. Volume 34. New York, New York, USA, (February 2020).

[26] Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38, 3, 50–57.

[27] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51, 5, 1–42.

[28] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A moral framework for understanding of fair ML through economic models of equality of opportunity. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (ACM FAT*)*. Atlanta, Georgia, USA, (January 2019).

[29] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency (ACM FAccT)*. Online, (March 2021). ISBN: 9781450383097.

[30] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133, 1, 237–293.

[31] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Leibniz International Proceedings in Informatics (LIPIcs)). Christos H. Papadimitriou, editor. Volume 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany. DOI: `10.4230/LIPIcs.ITCS.2017.43`. Retrieved 07/16/2019 from.

[32] Zachary C. Lipton. 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, 16, 3, 31–57.

[33] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Online, (July 2020).

[34] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22–31.

[35] Arvind Narayanan. 2018. Tutorial: 21 fairness definitions and their politics. FAT* 2018, (March 2018). `https://www.youtube.com/watch?v=jIXIuYdnyyk`.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[37] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (ACM FAT*)*. Association for Computing Machinery, New York, New York, USA, (January 2020). DOI: `10.1145/3351095.3372873`. Retrieved 01/19/2022 from.

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*. San Francisco, California, USA, (August 2016).

[39] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: research methods for detecting discrimination on internet platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a preconference at the 64th Annual Meeting of the International Communication Association*. Seattle, Washington, USA, (May 2014).

[40] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, (October 2017).

[41] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2019. A study in rashomon curves and volumes: a new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*.

[42] 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 (AIES)*. Online, (May 2021).

[43] Naeem Siddiqi. 2012. *Credit risk scorecards: developing and implementing intelligent credit scoring*. Volume 3. John Wiley & Sons.

[44] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. New York, New York, USA, (February 2020).

[45] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7, 2, 76–99.

[46] Rebecca Wexler. 2017. When a computer program keeps you in jail. *New York Times*. Retrieved January 15, 2022.

Broader Impact Statement

FRAUD-Detect may be used to provide evaluations of the fairness of model explanations, resulting in increased trust in explanations, public expectation of model fairness, and model owner accountability. Since FRAUD-Detect is non-cooperative and resists evasion even against an informed adversary, it empowers auditors, protects the intellectual property and ownership of the model owners, and may resist gamification into a new arms-race system between fairwashing and auditor detection methods. By empowering auditors, FRAUD-Detect also enforces democratic ideals of equality by providing an avenue for users to receive fairness-vetted model explanations for high-stakes decisions, further increasing public trust in AI and model auditing.

The potential risks of this work revolve around the fairness definition and datasets used in this paper. Demographic parity, while befitting our setting (as discussed in Section 2.1) is not always an appropriate fairness definition, depending on the model's domain and application. We recognise that continual focus on a concept reinforces it and it's ideology's normalcy, and that our focus on demographic parity also reinforces the political and philosophical ideologies underlying it and deprives alternate definitions of due attention. Likewise, as noted in Appendix E, our evaluation of FRAUD-Detect on the current popular fairness benchmark datasets normalizes these applications of AI. We draw specific attention to the use of the COMPAS recidivism prediction dataset [5], and note that we do not necessarily condone the use of AI models for this application.

To mitigate these societal risks, we advocate for similar auditing measures for second-order failures with respect to other fairness definitions (see Section 7). We also acknowledge the importance of investigating the impact of FRAUD-Detect in real-world scenarios, especially where the group of users suing for model explanations represents minority or sensitive demographics.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In the Societal Impact Statement
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix D
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See supplemental material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See supplemental material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See supplemental material.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Explanation methods

Methods to facilitate the understanding of machine learning models are usually classified into two main categories: transparent-box design and black-box model explanation [27]. Transparent-box design concerns methods that produce inherently interpretable models such as logistic regressors and decision trees [12, 32]. Black-box model explanation corresponds to approaches designed to explain how black-box models produce their outcomes. For instance, global explanations techniques (*e.g.*, TREPAN [20] and DECTEXT [11]) focus on explaining the whole logic of black-box models by training an inherently interpretable surrogate model. On the other hand, local explanations aim at explaining black-box models decisions on single data points (*e.g.*, LIME [38], and GRADCAM [40]). For simplicity, in this paper, we focus on fairwashing performed with global explanation models.

B Related Work

We highlight some of the differences in settings, assumptions and solutions between our work and [44, 25]:

Setting We rely on explainer models while they use a published data subset for fairness auditing. In particular, they reveal some subset of the training data and their predictions. We follow the setting considered in the literature introducing fairwashing [2, 3]. Slack et al. explanations require input perturbations while ours require the model owner to provide an interpretable model.

Assumptions Slack et al. assume that fairness auditing is performed via model-agnostic local explanations (e.g., LIME and SHAP) [44]. Both [44] and [25] assume an ideal detector with knowledge of the underlying training distribution of the model; and query access to the black-box model. However, we only assume access to black-box model predictions on the suing set. We stress that this dataset is available before any formal audit takes place (in the form of asking for model explanations). This does *not* constitute query access to the black-box model; in fact it is only dataset access, which here is the predictions on the suing set.

Solutions [44, 25] attempt to detect fairwashing using a Kolmogorov-Smirnov (KS) test over the underlying and company-provided subset distribution to determine if the data subset was honestly provided. More precisely, they show that detection is difficult when fairwashing is conducted by minimizing the Wasserstein distance between the distributions while subject to fairness; this optimization also minimizes the upper-bound of the advantage (i.e., distinguishability) of the KS test. This differs from our setting in which an informed malicious company must optimize the fairness subject to both fidelity (loss) and the detection threshold. Explanation methods like LIME and SHAP perturb inputs to gauge feature relevance, inadvertently querying with synthetic data that may be detected with out-of-distribution detection. Queries determined to originate from explainers are fed to a fair model whereas in-distribution points are given to the biased model. In contrast, our detection method is non-cooperative (we require no additional information from the black-box model as the auditor already has the predictions on the suing set) and therefore does not rely on input perturbations.

C Fidelity

Definition 6 (Fidelity [20])**.** *The fidelity of an interpretable model $I(\cdot)$ with respect to a black-box model $B(\cdot)$ is defined as the relative accuracy of $I(\cdot)$ with respect to $B(\cdot)$:*

$$\texttt{Fidelity}(I, B) = \Pr_{X \sim \mathfrak{D}}[\hat{Y}_B = \hat{Y}_I \mid X], \tag{10}$$

where $\mathfrak{D}$ is a data distribution, $\hat{Y}_I$ is the prediction of the interpretable model and $\hat{Y}_B$ is the prediction of black-box model. $\hat{Y}_I$ and $\hat{Y}_B$ are assumed to be both binary-valued random variables.

D Additional Proofs

Below we iterate the full proof for Theorem 1, the main theorem in the paper.

Proof. Define T_a^+, the true-positive rate of $I(\cdot)$ with respect to $B(\cdot)$ on $X \sim \mathfrak{D}_a$, in which $\mathfrak{D}_a$ denotes the distribution over data with attribute $a \in A$. Define F_a^+, the false-positive rate of $I(\cdot)$ with respect to $B(\cdot)$ on $X \sim \mathfrak{D}_a$:

$$T_a^+ = \Pr[\hat{Y}_I = 1 \mid \hat{Y}_B = 1, A = a], \quad F_a^+ = \Pr[\hat{Y}_I = 1 \mid \hat{Y}_B = 0, A = a] \qquad \text{For} \quad a \in \{0, 1\}. \tag{11}$$

Note that by definition, $T_a^+, F_a^+ \in [0, 1]$. We denote $\tilde{\delta}, \delta' > 0$ as the difference in value between T_0^+ and T_1^+, and F_0^+ and F_1^+, respectively. Without loss of generality, let $T_1^+ = T_0^+ + \tilde{\delta}$ and $F_1^+ = F_0^+ + \delta'$ where $\tilde{\delta}, \delta' \in [0, 1]$.

To characterize $\gamma = \Gamma_B - \Gamma_I$, we first express Γ_I and Γ_B in terms of $\tilde{\delta}$, δ', T_0^+, T_1^+, F_0^+, and F_1^+, and $Y_B|A$. Recall from Definition 3 that we define $\Gamma_I := \Pr\left[\hat{Y}_I = 1 \mid A = 0\right] - \Pr\left[\hat{Y}_I = 1 \mid A = 1\right]$.

We use Bayes' formula to re-write $\Pr\left[\hat{Y}_I = 1 \mid A = a\right]$, $a \in A = \{0, 1\}$ in terms of the aforementioned variables. For $a \in \{0, 1\}$:

$$\begin{aligned}\Pr\left[\hat{Y}_I = 1 \mid A = a\right] &= \Pr\left[\hat{Y}_B = 1 \mid A = a\right] \Pr\left[\hat{Y}_I = 1 \mid \hat{Y}_B = 1, A = a\right] \\ &\quad + \Pr\left[\hat{Y}_B = 0 \mid A = a\right] \Pr\left[\hat{Y}_I = 1 \mid \hat{Y}_B = 0, A = a\right] \\ &= \Pr\left[\hat{Y}_B = 1 \mid A = a\right] \cdot T_a^+ + \Pr\left[\hat{Y}_B = 0 \mid A = a\right] \cdot F_a^+ \end{aligned} \tag{12}$$

Given our derivation above, we continue by substituting terms into Γ_I:

$$\begin{aligned}\Gamma_I &= \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] \cdot T_0^+ + \Pr\left[\hat{Y}_B = 0 \mid A = 0\right] \cdot F_0^+ \\ &\quad - \Pr\left[\hat{Y}_B = 1 \mid A = 1\right] \cdot T_1^+ - \Pr\left[\hat{Y}_B = 0 \mid A = 1\right] \cdot F_1^+ \\ &= \left(T_1^+ + \tilde{\delta}\right) \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] + \left(F_1^+ + \delta'\right) \Pr\left[\hat{Y}_B = 0 \mid A = 0\right] \\ &\quad - T_1^+ \Pr\left[\hat{Y}_B = 1 \mid A = 1\right] - F_1^+ \Pr\left[\hat{Y}_B = 0 \mid A = 1\right] \\ &= T_1^+ \left(\Pr\left[\hat{Y}_B = 1 \mid A = 0\right] - \Pr\left[\hat{Y}_B = 1 \mid A = 1\right]\right) \\ &\quad + F_1^+ \left(\Pr\left[\hat{Y}_B = 0 \mid A = 0\right] - \Pr\left[\hat{Y}_B = 0 \mid A = 1\right]\right) \\ &\quad + \tilde{\delta} \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] + \delta' \Pr\left[\hat{Y}_B = 0 \mid A = 0\right] \\ &= T_1^+ \Gamma_B - F_1^+ \Gamma_B + \tilde{\delta} \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] + \delta' \left(1 - \Pr\left[\hat{Y}_B = 1 \mid A = 0\right]\right) \\ &= \Gamma_B \left(T_1^+ - F_1^+\right) + \left(\tilde{\delta} - \delta'\right) \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] + \delta'\end{aligned}$$

Now that we have derived Γ_I in terms of our desired terms and Γ_B, we eliminate Γ_I from the equation for γ:

$$\begin{aligned}\gamma &= \Gamma_B - \Gamma_I \\ &= \Gamma_B - (\Gamma_B(T_1^+ - F_1^+) + (\tilde{\delta} - \delta') \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] + \delta') \\ &= \Gamma_B(1 + F_1^+ - T_1^+) + (\delta' - \tilde{\delta}) \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] - \delta' \\ &= \Gamma_B(1 + F_1^+ - T_1^+) - \delta' \cdot \left(1 - \Pr\left[\hat{Y}_B = 1 \mid A = 0\right]\right) - \tilde{\delta} \Pr\left[\hat{Y}_B = 1 \mid A = 0\right] \\ &= \Gamma_B(1 + F_1^+ - T_1^+) - \delta' \Pr\left[\hat{Y}_B = 0 \mid A = 0\right] - \tilde{\delta} \Pr\left[\hat{Y}_B = 1 \mid A = 0\right].\end{aligned} \tag{13}$$

□

Corollary 1. *Let T_1^+, F_1^+ as defined in Section 3. Then:*

$$1 + F_1^+ - T_1^+ > 0.$$

Proof. Note that $F_1^+, T_1^+ \in [0, 1]$, therefore $1 + F_1^+ - T_1^+ \geq 0$. We will show $1 + F_1^+ - T_1^+ = 0$ if and only if $T^+ = 1$ and $F^+ = 0$, *i.e.* if $I(\cdot)$ has perfect fidelity w.r.t. $B(\cdot)$. We re-arrange $1 + F_1^+ - T_1^+ = 0$ as:

$$T_1^+ - F_1^+ = 1$$

Since $F_1^+, T_1^+ \in [0, 1]$ and the size of interval $[0, 1]$ is $1 - 0 = 1 = T_1^+ - F_1^+$, T_1^+ and F_1^+ must be endpoints 1 and 0, respectively. □

E Details on Experimental Setup of FRAUD-Detect

In this section, we provide additional details on our experimental setup.

Table 1: The accuracy (Acc.) and demographic parity gap (Gap) of the black-box models–AdaBoost, Deep Neural Networks (DNN), Random Forests (RF) and Gradient Boosted Decision Trees (XgBoost) evaluated on the suing set.

Dataset	AdaBoost		DNN		RF		XgBoost	
	Acc.	Gap	Acc.	Gap	Acc.	Gap	Acc.	Gap
COMPAS	0.68	0.21	0.68	0.25	0.67	0.25	0.68	0.27
Adult Income	0.85	0.12	0.85	0.16	0.86	0.17	0.86	0.17
Bank Marketing	0.91	0.07	0.91	0.09	0.91	0.10	0.91	0.10

Datasets. COMPAS [5] is a recidivism dataset released by ProPublica that includes data on $6,167$ offenders in prison from Broward County (Florida), with the classification task consisting in predicting whether an inmate will re-offend in the two years after their release.[2] We consider race (`African-American`, `Caucasian`) as the sensitive attribute in COMPAS. Adult Income [22] includes 48,842 profiles of individuals, each characterized by 13 attributes, drawn from the U.S census. The Adult Income binary classification task is to predict whether an individual's income will exceed $50,000 a year. We use gender (`Female`, `Male`) as the sensitive attribute in Adult Income. Finally, Bank Marketing [34] constitutes information collected from a direct marketing campaign of a Portuguese banking institution between 2008 to 2013. This dataset contains 41,188 profiles of individuals, each described by 20 attributes with the classification task being the prediction of subscription to a term deposit. We consider age (`between 30-60` or `not between 30-60`) as the sensitive attribute in Bank Marketing.

Training of black-box and interpretable models. For each experiment, the dataset is split randomly into three partitions: the training set (67%), the suing set (16.5%) and the testing set (16.5%), following the same procedure as in [3]. Each experiment is repeated with 10 separate random seeds. The black-box models are learned from the training data points with their associated true labels. However, the interpretable models are trained from the suing set, which has been labelled by the black-box model. More precisely, we use the exponentiated gradient algorithm [1] subject to the fairwashing constraint from the FairLearn Library [9] to learn the fairwashed interpretable model. The exponentiated gradient algorithm when applied for fairness identifies the saddle point solution jointly maximizing the fairness and minimizing the loss. To learn the fairwashed interpretable model of the informed adversary, we extend the `Fairness In The Rashomon Set` (FaiRS) [19] framework by adding an additional constraint on $\mathcal{C}_{\text{KL}}$. FaiRS reduces the problem of minimizing the unfairness under two constraints on the fidelity and $\mathcal{C}_{\text{KL}}$ to a saddle point problem, which is solved with the exponentiated gradient technique [1]. Table 1 summarizes the performance of the considered black-box models on each suing set.

Implementation. FRAUD-Detect solely requires the confusion matrices from the interpretable model for both the majority and minority group. These matrices are flattened, then compared via KL as outlined in Equation 8. Experiments were conducted on Intel Xeon Silver 4110 CPUs (2.10 GHz, 8GB RAM) and the KL divergence detector implemented in Python 3.8 using scikit-learn [36]. Given that the KL divergence constraint is not a linear constraint on the confusion matrices, the extended version of `FaiRS` may fail to obtain the global minimum of the demographic parity gap on the set of high-fidelity interpretable models. In our experiments, our approach converges for two datasets, namely COMPAS and Marketing, but does not achieve the global minimum of the demographic parity gap for Adult Income. We are planning to address this in future work by focusing on detecting fairwashing using linear constraints.

[2] We note that considerable bias and confounding exists in the true labels for recidivism datasets at large. We do not necessarily condone use of recidivism datasets. However, COMPAS exists as a standard fairness benchmark and thus, we use COMPAS to validate our method.

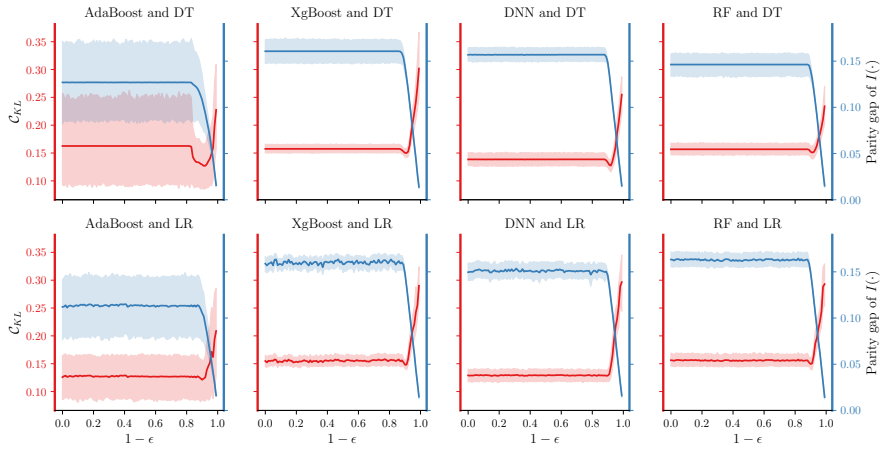

Figure 4: Kullback–Leibler divergence between confusion matrices of subpopulations, $\mathcal{C}_{\text{KL}}$ and demographic parity gap as a function of ϵ fairwashing in interpretable models explaining black-box models using Adult Income dataset. We use Logistic Regression (LR) and Decision Tree (DT) as interpretable models, while AdaBoost, Deep Neural Network (DNN), Random Forest (RF) and Gradient Boosted Decision Trees (XgBoost) are used as black-box models.

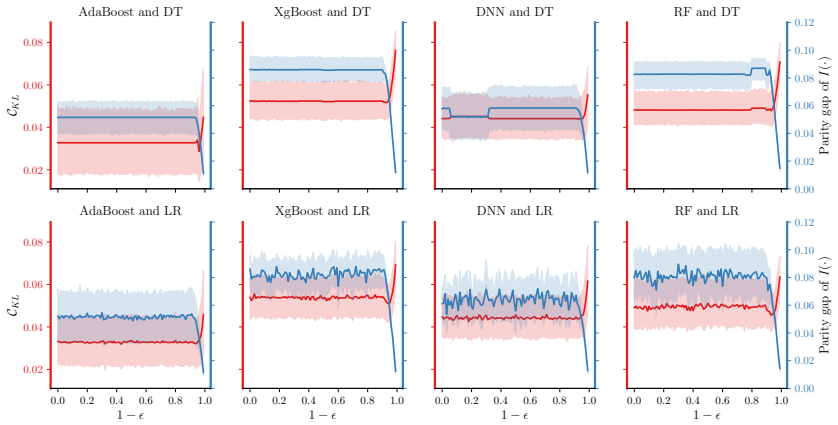

Figure 5: Kullback–Leibler divergence between confusion matrices of subpopulations, $\mathcal{C}_{\text{KL}}$ and demographic parity gap as a function of ϵ fairwashing in interpretable models explaining black-box models using Marketing dataset. We use Logistic Regression (LR) and Decision Tree (DT) as interpretable models, while AdaBoost, Deep Neural Network (DNN), Random Forest (RF) and Gradient Boosted Decision Trees (XgBoost) are used as black-box models.

F Additional results

FRAUD-Detect successfully detects fairwashing. Figures 4 and 5 illustrate both the demographic parity gap and $\mathcal{C}_{\text{KL}}$ of interpretable models as a function of ϵ for two datasets (Adult income and Bank marketing), respectively.

FRAUD-Detect is robust to the informed adversary. Figures 6 and 7 show the range of the demographic parity gap of high-fidelity fairwashed interpretable models subjected to a constraint on the $\mathcal{C}_{\text{KL}}$ by solving the optimization problem of Equation (9). We use four types of black-box models (AdaBoost, XgBoost, DNN and Random Forest).

G Fairwashing Detection Thresholds

Table 2 reports several fairwashing detection thresholds observed in Figures 2, 4 and 5.

Table 2: $\mathcal{C}_{KL}$ values for each dataset, black-box model, and interpretable model. These values are chosen from average KL divergence values at 5% and 50% fairwashing.

Dataset	IM	AdaBoost 5%	AdaBoost 50%	DNN 5%	DNN 50%
COMPAS	DT	0.105 ± 0.035	0.206 ± 0.078	0.136 ± 0.037	0.185 ± 0.065
	LR	0.101 ± 0.029	0.164 ± 0.075	0.143 ± 0.044	0.231 ± 0.104
Adult Income	DT	0.128 ± 0.067	0.148 ± 0.086	0.134 ± 0.015	0.154 ± 0.029
	LR	0.132 ± 0.069	0.170 ± 0.117	0.128 ± 0.017	0.213 ± 0.048
Bank Marketing	DT	0.035 ± 0.022	0.034 ± 0.025	0.045 ± 0.017	0.046 ± 0.018
	LR	0.033 ± 0.018	0.036 ± 0.022	0.042 ± 0.015	0.049 ± 0.020
Dataset	**IM**	**RF 5%**	**RF 50%**	**XgBoost 5%**	**XgBoost 50%**
COMPAS	DT	0.149 ± 0.040	0.227 ± 0.095	0.170 ± 0.041	0.248 ± 0.119
	LR	0.145 ± 0.043	0.215 ± 0.088	0.164 ± 0.044	0.261 ± 0.124
Adult Income	DT	0.154 ± 0.017	0.170 ± 0.024	0.153 ± 0.011	0.196 ± 0.036
	LR	0.151 ± 0.016	0.207 ± 0.044	0.148 ± 0.012	0.200 ± 0.041
Bank Marketing	DT	0.048 ± 0.013	0.054 ± 0.014	0.053 ± 0.015	0.059 ± 0.016
	LR	0.045 ± 0.014	0.048 ± 0.015	0.052 ± 0.014	0.053 ± 0.015

H Per seed fine grain C_{KL} versus fairness

Figures 8, 9, and 10 show the per-seed results aggregated to form figures 2, 4, and 5, respectively. In these figures (especially in 8 and 10), the order of increasing black-box fairnesses across the 10 seeds is often replicated in the order of decreasing interpretable model $\mathcal{C}_{\text{KL}}$ values across the same seeds; thus a seed with a low black-box fairness has higher $\mathcal{C}_{\text{KL}}$ values than a seed with high black-box fairness.

Some seeds from the logistic regression interpretable model (e.g. Figure 9 AdaBoost and logistic regression green and grey lines) show discontinuities at high epsilons where fairwashing is significant. There, the $\mathcal{C}_{\text{KL}}$ values jump to infinity due to a division by zero in the KL log term when there are 0s in the minority subpopulation confusion matrix.

We also observe that for several seeds across all datasets and models, the fairness of the honest interpretable model is imperfectly aligned and usually higher than the black-box fairness. This is unintentional fairwashing, which may arise from the relatively small capacity of the interpretable model compared to the black-box model.

I Bounds for KL Divergence of the Confusion Matrices

First, denote $\gamma_1 = \Gamma_B(1 + F_1^+ - T_1^+) = \Gamma_B(F_1^- + F_1^+)$ and similarly $\gamma_0 = \Gamma_B(F_0^- + F_0^+)$. Depending on the context of the problem either subgroup could be the more sensitive subgroup, so in Equation (6) we implicitly assumed that subgroup is $z = 1$: $\gamma = \gamma_1$. Equation (8) defines the detector based on the flattened confusion matrix, but since for a binary classification problem, the rank of the confusion matrix is 2, it is enough to consider $\text{KL}\left(\begin{bmatrix} F_0^+ \\ F_0^- \end{bmatrix}, \begin{bmatrix} F_1^+ \\ F_1^- \end{bmatrix}\right)$.[3] We will derive upper and lower bounds on this divergence measure in terms of $\gamma_0, \gamma_1, F_i^+, F_i^-;\ i \in \{0, 1\}$. First,

[3] That is to say, any two rates from the confusion matrix is enough to construct the full matrix. We note that in this section, we work with false positive rate F^+ and false negative rate F^- to simplify the presentation. The rest of the paper uses the entire confusion matrix which as we have pointed out in Theorem 1 is not necessary but is also harmless.

note that γ_1 is the sum of type I F_1^- and type II F_1^+ detection errors; scaled by the unfairness of the black-box model Γ_B. In a detection problem, typically, we bound the type I error and minimize the type II error. The particular choice of type I and type II error is context-dependent (are false positives a bigger concern than false negatives or not?). In the context of fairwashing, a reasonable choice is that

$$\begin{aligned} F_1^+ &\leq F_1^-, \\ F_0^+ &\leq F_0^- \end{aligned} \tag{14}$$

That is, the cost of falsely detecting a fairwashed model is larger than missing fairwashing—due to the legal and liability considerations, for instance.

We can bound the irreducible fairwashing gap γ_i:

$$\begin{aligned} 2\Gamma_b F_1^+ &\leq \gamma_1 \leq 2\Gamma_b F_1^-, \\ 2\Gamma_b F_0^+ &\leq \gamma_0 \leq 2\Gamma_b F_0^- \end{aligned} \tag{15}$$

Using this we bound the aforementioned KL divergence:

$$\begin{aligned} \mathrm{KL}\left(\begin{bmatrix} F_0^+ \\ F_0^- \end{bmatrix}, \begin{bmatrix} F_1^+ \\ F_1^- \end{bmatrix}\right) &= F_0^+ \log\left(\frac{F_0^+}{F_1^+}\right) + F_0^- \log\left(\frac{F_0^-}{F_1^-}\right) && (16) \\ &\leq F_0^- \log\left(\frac{F_0^+}{F_1^+} \cdot \frac{F_0^-}{F_1^-}\right) && (17) \\ &= F_0^- \log\left(\frac{F_0^+}{F_1^-} \cdot \frac{F_0^-}{F_1^+}\right) && (18) \\ &\leq F_0^- \log\left(\frac{\gamma_0/2\Gamma_B}{\gamma_1/2\Gamma_B} \cdot \frac{F_0^-}{F_1^+}\right) && (19) \\ &= F_0^- \log\left(\frac{\gamma_0}{\gamma_1} \cdot \frac{F_0^-}{F_1^+}\right) && (20) \end{aligned}$$

Similarly,

$$\begin{aligned} \mathrm{KL}\left(\begin{bmatrix} F_0^+ \\ F_0^- \end{bmatrix}, \begin{bmatrix} F_1^+ \\ F_1^- \end{bmatrix}\right) &\geq F_0^+ \log\left(\frac{F_0^+}{F_1^+} \cdot \frac{F_0^-}{F_1^-}\right) && (21) \\ &= F_0^+ \log\left(\frac{F_0^+}{F_1^-} \cdot \frac{F_0^-}{F_1^+}\right) && (22) \\ &\geq F_0^+ \log\left(\frac{F_0^+}{F_1^-} \cdot \frac{\gamma_0/2\Gamma_B}{\gamma_1/2\Gamma_B}\right) && (23) \\ &= F_0^+ \log\left(\frac{F_0^+}{F_1^-} \cdot \frac{\gamma_0}{\gamma_1}\right) && (24) \end{aligned}$$

Therefore, $\kappa := \mathrm{KL}\left(\begin{bmatrix} F_0^+ \\ F_0^- \end{bmatrix}, \begin{bmatrix} F_1^+ \\ F_1^- \end{bmatrix}\right)$ is bounded from below and above by:

$$F_0^+ \log\left(\frac{F_0^+}{F_1^-} \cdot \frac{\gamma_0}{\gamma_1}\right) \leq \kappa \leq F_0^- \log\left(\frac{\gamma_0}{\gamma_1} \cdot \frac{F_0^-}{F_1^+}\right) \tag{25}$$

Equivalently, we can re-write these bounds in terms of F^+ and T^+ to match Theorem 1 (using the fact that $F^- = 1 - T^+$):

$$F_0^+ \log\left(\frac{F_0^+}{1 - T_0^+} \cdot \frac{\gamma_0}{\gamma_1}\right) \leq \kappa \leq (1 - T_0^+) \log\left(\frac{\gamma_0}{\gamma_1} \cdot \frac{1 - T_0^+}{F_1^+}\right) \tag{26}$$

Therefore any choice between $\kappa_{min} := F_0^+ \log\left(\frac{F_0^+}{F_1^-} \cdot \frac{\gamma_0}{\gamma_1}\right)$ and $\kappa_{max} := F_0^- \log\left(\frac{\gamma_0}{\gamma_1} \cdot \frac{F_0^-}{F_1^+}\right)$ is valid. $\kappa = \kappa_{min}$ is the most conservative choice which increases the risk of false discovery (of fairwashing),

while $\kappa = \kappa_{max}$ is a permissive choice that minimizes the risk of false discovery at the expense of increase in the false negatives (not detecting fairwashing). A balanced choice for threshold could be the intermediate point:

$$\bar{\kappa} = \frac{1}{2}\left(F_0^+ \log\left(\frac{F_0^+}{F_1^-}\cdot\frac{\gamma_0}{\gamma_1}\right) + F_0^- \log\left(\frac{\gamma_0}{\gamma_1}\cdot\frac{F_0^-}{F_1^+}\right)\right) \tag{27}$$

$$= \frac{1}{2}\log\left(\left(\frac{F_0^+}{F_1^-}\right)^{F_0^+}\left(\frac{F_0^-}{F_1^+}\right)^{F_0^-}\left(\frac{\gamma_0}{\gamma_1}\right)^{F_0^+ + F_0^-}\right) \tag{28}$$

$$= \frac{1}{2}\log\left(\left(\frac{F_0^+}{F_1^-}\right)^{F_0^+}\left(\frac{F_0^-}{F_1^+}\right)^{F_0^-}\left(\frac{\gamma_0}{\gamma_1}\right)^{\gamma_0/\Gamma_B}\right) \tag{29}$$

$$= \frac{1}{2}\log\left(\left(\frac{F_0^+}{F_1^-}\right)^{F_0^+}\left(\frac{F_0^-}{F_1^+}\right)^{F_0^+ - \gamma_0/\Gamma_B}\left(\frac{\gamma_0}{\gamma_1}\right)^{\gamma_0/\Gamma_B}\right) \tag{30}$$

$$= \frac{1}{2}\log\left(\left(\frac{F_0^+F_0^-}{F_1^-F_1^+}\right)^{F_0^+}\left(\frac{F_0^-}{F_1^+}\right)^{-\gamma_0/\Gamma_B}\left(\frac{\gamma_0}{\gamma_1}\right)^{\gamma_0/\Gamma_B}\right) \tag{31}$$

$$= \frac{1}{2}\log\left(\left(\frac{F_0^+}{F_1^+}\right)^{F_0^+}\left(\frac{F_0^-}{F_1^-}\right)^{F_0^+}\left(\frac{F_0^-}{F_1^+}\right)^{-\gamma_0/\Gamma_B}\left(\frac{\gamma_0}{\gamma_1}\right)^{\gamma_0/\Gamma_B}\right) \tag{32}$$

$$= \frac{1}{2}\log\left(\left(\frac{F_0^+}{F_1^+}\right)^{F_0^+}\left(\frac{F_0^-}{F_1^-}\right)^{F_0^+}\left(\frac{\gamma_1}{\Gamma_B} - \frac{F_1^-}{F_0^-}\right)^{\gamma_0/\Gamma_B}\left(\frac{\gamma_0}{\gamma_1}\right)^{\gamma_0/\Gamma_B}\right) \tag{33}$$

$$= \frac{F_0}{2}\left(\Delta_l(F^+) + \Delta_l(F^-)\right) + \frac{\gamma_0}{2\Gamma_B}\log\left(\frac{\gamma_0}{\Gamma_B} - \frac{\gamma_0}{\gamma_1}\frac{F_1^-}{F_0^-}\right) \tag{34}$$

$$= \frac{F_0}{2}\left(\Delta_l(F^+) + \Delta_l(F^-)\right) + \frac{\gamma_0}{2\Gamma_B}\log\left(\frac{\gamma_0}{\Gamma_B}\right) + \frac{\gamma_0}{2\Gamma_B}\log\left(1 - \frac{\Gamma_B}{\gamma_1}\frac{F_1^-}{F_0^-}\right) \tag{35}$$

$$= \frac{F_0}{2}\left(\Delta_l(F^+) + \Delta_l(F^-)\right) + \frac{\gamma_0}{2\Gamma_B}\log\left(\frac{\gamma_0}{\Gamma_B}\right) - \frac{\gamma_0}{2\Gamma_B}\left(\frac{\Gamma_B}{\gamma_1}\frac{F_1^-}{F_0^-}\right) - O\left(\frac{\Gamma_B^2}{\gamma_1^2}\left(\frac{F_1^-}{F_0^-}\right)^2\right) \tag{36}$$

$$= \frac{F_0}{2}\left(\Delta_l(F^+) + \Delta_l(F^-)\right) + \frac{\gamma_0}{2\Gamma_B}\log\left(\frac{\gamma_0}{\Gamma_B}\right) - \frac{1}{2}\frac{\gamma_0}{\gamma_1}\left(\frac{F_1^-}{F_0^-}\right) - O\left(\frac{\Gamma_B^2}{\gamma_1^2}\left(\frac{F_1^-}{F_0^-}\right)^2\right) \tag{37}$$

$$\bar{\kappa} = \frac{F_0}{2}\left(\Delta_l(F^+) + \Delta_l(F^-)\right) + \frac{\gamma_0}{2\Gamma_B}\log\left(\frac{\gamma_0}{\Gamma_B}\right) - \frac{1}{2}\frac{\gamma_0}{\gamma_1}\frac{1}{\Delta(F^-)} - O\left(\frac{\Gamma_B^2}{\gamma_1^2}\left(\frac{F_1^-}{F_0^-}\right)^2\right) \tag{38}$$

where $\Delta(R) = R_0/R_1$ and $\Delta_l(R) = \log\Delta(R)$. In the last equation, we have used the taylor expansion of $\log(1-x)$ at $x = 0$.

J Experimental results with two independent variables

In this section, we repeat experiments shown in Section 6 with the following adjustment:

$$\mathcal{C}_{KL} = \mathrm{KL}([T_0^+, F_0^+], [T_1^+, F_1^+]),$$

in which we use only two independent variables of true-positive rate T^+ and false-positive rate F^+. Figures 11, 12 and 13 show similar results as Figures 2, 4 and 5.

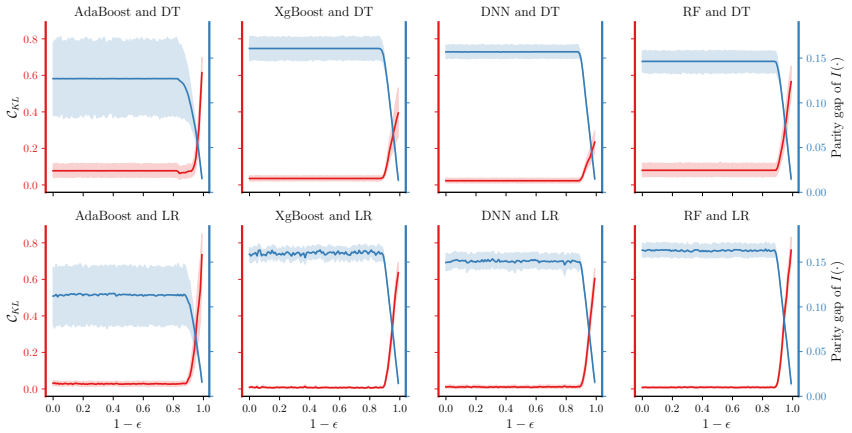

Figure 12: KL among altered confusion matrices of subgroups versus demographic parity unfairness of two explainable models, namely Logistic Regression (LR) and Decision Tree (DT), approximating four black-box models, namely AdaBoost, Deep Neural Network (DNN), Random Forest (RF), and Gradient Boosted Decision Trees (XgBoost) using Adult dataset. We consider 10 split of the dataset to the training set for the black-box model and suing set for the interpretable model.

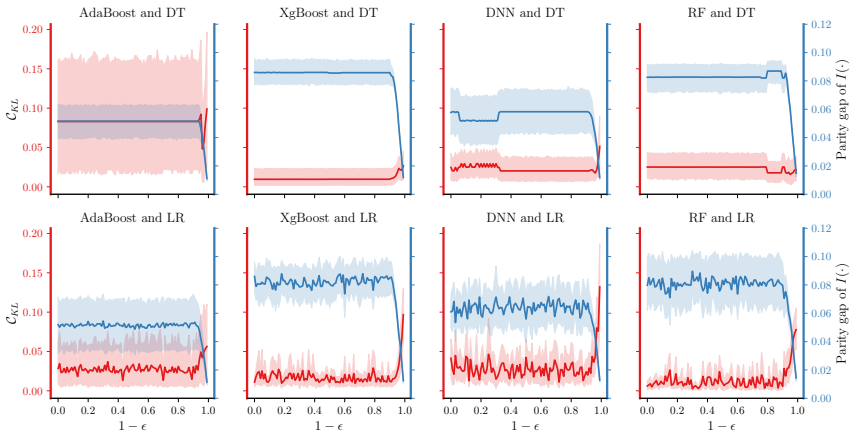

Figure 13: KL among altered confusion matrices of subgroups versus demographic parity unfairness of two explainable models, namely Logistic Regression (LR) and Decision Tree (DT), approximating four black-box models, namely AdaBoost, Deep Neural Network (DNN), Random Forest (RF), and Gradient Boosted Decision Trees (XgBoost), using Marketing dataset. We consider 10 split of the dataset to the training set for the black-box model and suing set for the interpretable model.

K Perfect Fairwashing is impossible

Theorem 2. *Consider binary-valued random variables $\hat{Y}_B$ black-box model predictions, $\hat{Y}_I$ proxy (interpretable) model predictions, and A a sensitive attribute; then these conditions cannot be satisfied simultaneously:*

1. *Interpretable Model Demographic Parity*

$$\Pr\left[\hat{Y}_I = 1 \mid A = 0\right] = \left[\hat{Y}_I = 1 \mid A = 1\right]$$

,

2. Non-detectable

$$\Pr\left[Y_I = 1 \mid Y_B = 1, A = 0\right] = \Pr\left[Y_I = 1 \mid Y_B = 1, A = 1\right] := T^+ \tag{39}$$

$$\Pr\left[Y_I = 1 \mid Y_B = 0, A = 0\right] = \Pr\left[Y_I = 1 \mid Y_B = 0, A = 1\right] := F^+. \tag{40}$$

Proof. Let Equation (39) be equal to T^+, and Equation (40) to F^+. Define the base rates of the black-box model on subpopulations 0 and 1 as:

$$\mu_0 = \Pr[\hat{Y}_B = 1 | A = 0]$$
$$\mu_1 = \Pr[\hat{Y}_B = 1 | A = 1].$$

Using the rule of total probabilities, for $a \in [0, 1]$,

$$\begin{aligned}\Pr\left[\hat{Y}_I = 1 \mid A = a\right] &= \Pr[\hat{Y}_B = 1 | A = a] \Pr\left[\hat{Y}_I = 1 \mid \hat{Y}_B = 1, A = a\right] \\ &\quad + \Pr[\hat{Y}_B = 0 | A = a] \Pr\left[\hat{Y}_I = 1 \mid \hat{Y}_B = 0, A = a\right] \qquad (41) \\ &= \mu_a \cdot T^+ + (1 - \mu_a) \cdot F^+ \qquad (42)\end{aligned}$$

Fairness according to parity requires that

$$\Pr\left[\hat{Y}_I = 1 \mid A = 0\right] \stackrel{?}{=} \Pr\left[\hat{Y}_I = 1 \mid A = 1\right]. \tag{43}$$

Substituting Equation (41) and Equation (43) and re-organizing we have:

$$(\mu_0 - \mu_1) \cdot T^+ + ((1 - \mu_0) - (1 - \mu_1)) \cdot F^+ \stackrel{?}{=} 0 \tag{44}$$

Case 1: If $T^+ = F^+$, by eliminating both and recognizing that $(1 - \mu_0) = 1 - \mu_0$, we see that the equality Equation (43) reduces to $0 = 0$ which is trivially correct.

Case 2: $T^+ \neq F^+$ Simplifying Equation (44) further and eliminating $T^+ - F^+ \neq 0$

$$\begin{aligned}&\mu_0(T^+ - F^+) - \mu_1 \cdot (T^+ - F^+) \stackrel{?}{=} (T^+ - F^+) \\ \Leftrightarrow \quad &\mu_0 \stackrel{?}{=} 1 + \mu_1.\end{aligned}$$

But these are probabilities, for Equation (43) to hold we must have that

$$\begin{aligned}\mu_1 &= 0 \\ \mu_0 &= 1,\end{aligned} \tag{45}$$

which means that the black-box predictions exclusively depend on the sensitive attribute. In other words, for demographic parity to hold for the interpretable model, the black-box models predictions all samples of one subpopulation $A = 0$ and reject the others $A = 0$. Clearly, this is unfair to one subpopulation.

To establish the impossibility result, we further show that such an assumption is equivalent to requiring completely inaccurate black-box models with maximum error rates.

Consider the "true" base rates for each subpopulation:

$$\Pr[Y = 1 \mid A = 0] = \Pr\left[Y = 1 \mid \hat{Y}_B = 1, A = 0\right] \mu_0 + \Pr\left[Y = 1 \mid \hat{Y}_B = 0, A = 0\right] \Pr\left[\hat{Y}_B = 0 \mid A = 0\right]$$
$$\Pr[Y = 1 \mid A = 1] = \Pr[Y = 1 \mid \hat{Y}_B = 1, A = 1]\mu_1 + \Pr\left[Y = 1 \mid \hat{Y}_B = 1, A = 1\right] (1 - \mu_1)$$

If interpretable demographic parity Equation (45) holds then, these base rates reduce to:

$$\Pr[Y = 1 \mid A = 0] = \Pr\left[Y = 1 \mid \hat{Y}_B = 1, A = 0\right]$$
$$\Pr[Y = 1 \mid A = 1] = \Pr\left[Y = 1 \mid \hat{Y}_B = 0, A = 1\right] \Longrightarrow Y \perp \hat{Y}_B \mid A \Longrightarrow \hat{Y}_B \perp Y \mid A.$$

In other words, the black-box model predictions is completely independent of the true label and exclusively dependent on the sensitive attribute A. □

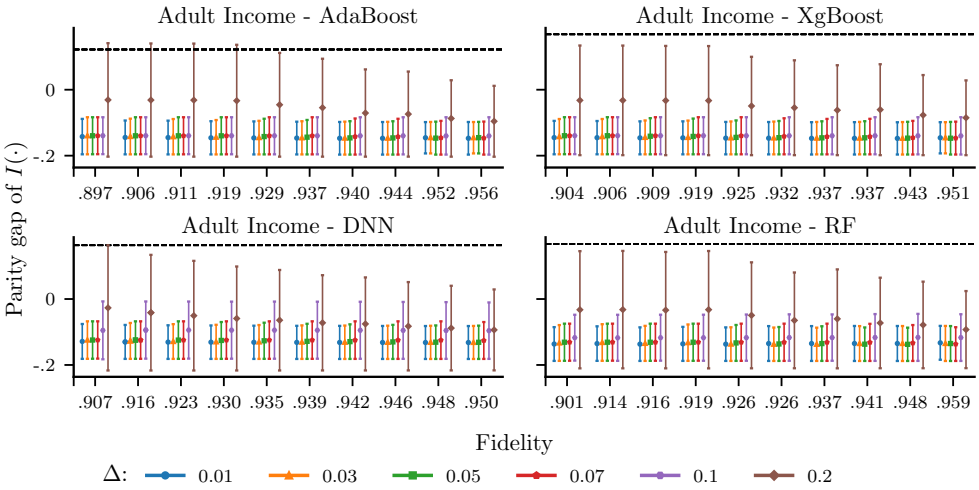

Figure 6: Range of the demographic parity gap of high-fidelity fairwashed logistic regression interpretable models subjected to a constraint on the $\mathcal{C}_{\text{KL}}$ ($\Delta \in \{0.01, 0.03, 0.05, 0.07, 0.10, 0.20\}$), explaining AdaBoost, Deep Neural Network (DNN), Random Forest (RF) and Gradient Boosted Decision Trees (XgBoost) black-box models trained on Adult Income dataset. Horizontal lines denote the demographic parity gap of the black-box models.

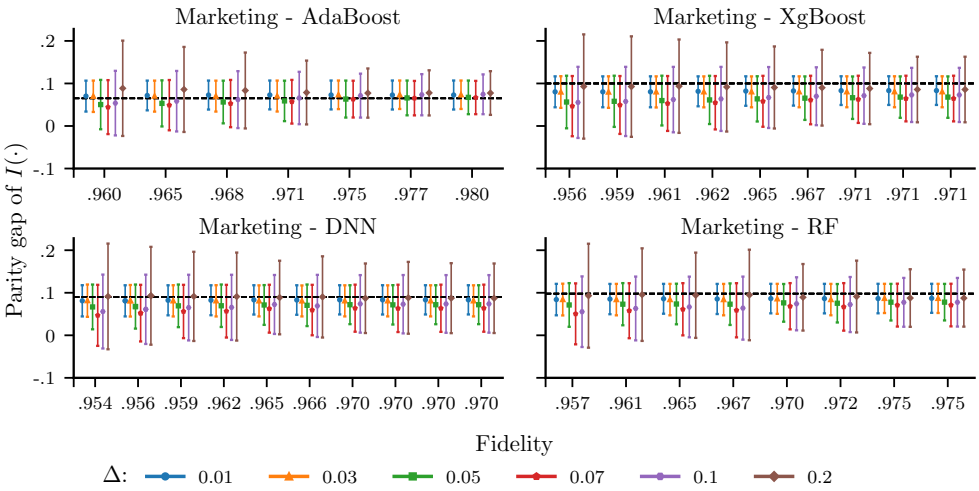

Figure 7: Range of the demographic parity gap of high-fidelity fairwashed logistic regression interpretable models subjected to a constraint on the $\mathcal{C}_{\text{KL}}$ ($\Delta \in \{0.01, 0.03, 0.05, 0.07, 0.10, 0.20\}$), explaining AdaBoost, Deep Neural Network (DNN), Random Forest (RF) and Gradient Boosted Decision Trees (XgBoost) black-box models trained on Marketing dataset. Horizontal lines denote the demographic parity gap of the black-box models.

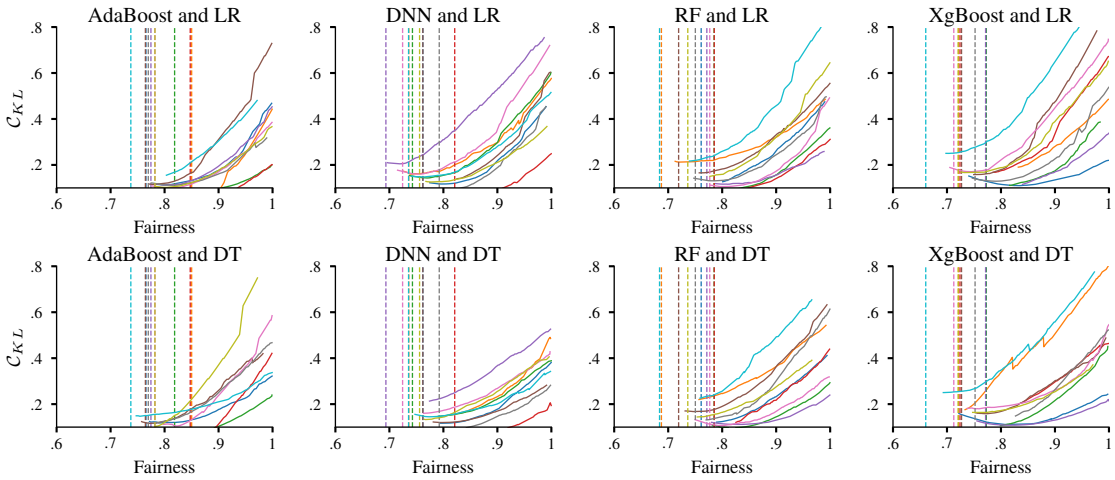

Figure 8: Kullback–Leibler divergence, $\mathcal{C}_{\text{KL}}$, between confusion matrices of subgroups versus demographic parity fairness of two interpretable models, namely Logistic Regression (LR) and Decision Tree (DT), approximating four black-box models, namely AdaBoost, Deep Neural Network (DNN), Random Forest (RF), and Gradient Boosted Decision Trees (XgBoost), using COMPAS dataset. We consider 10 split of the dataset to the training set for the black-box model and suing set for the interpretable model. Vertical dashed lines represent unfairness values of the black-box models.

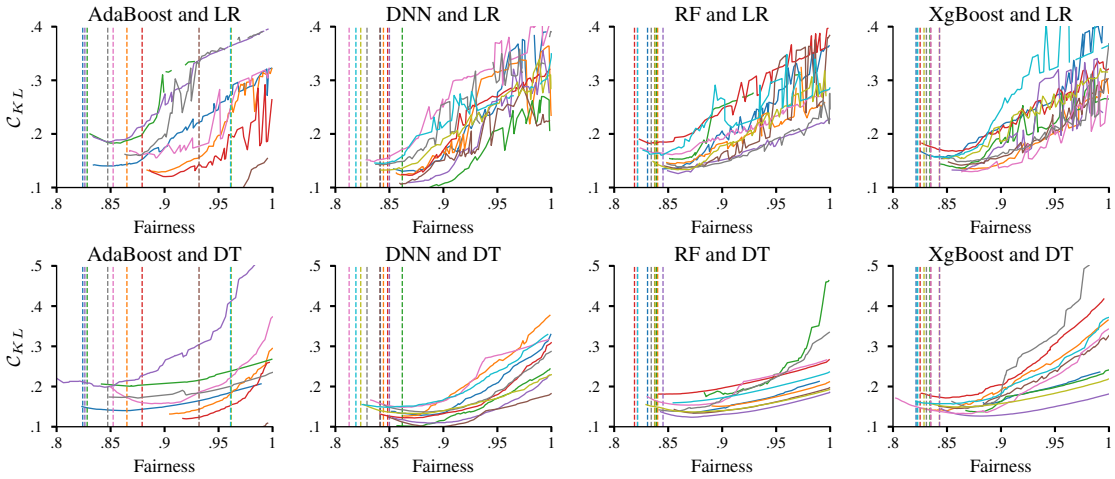

Figure 9: Kullback–Leibler divergence, $\mathcal{C}_{\text{KL}}$, between confusion matrices of subgroups versus demographic parity fairness of two interpretable models, namely Logistic Regression (LR) and Decision Tree (DT), approximating four black-box models, namely AdaBoost, Deep Neural Network (DNN), Random Forest (RF), and Gradient Boosted Decision Trees (XgBoost) using Adult dataset. We consider 10 split of the dataset to the training set for the black-box model and suing set for the interpretable model. Vertical dashed lines represent unfairness values of the black-box models.

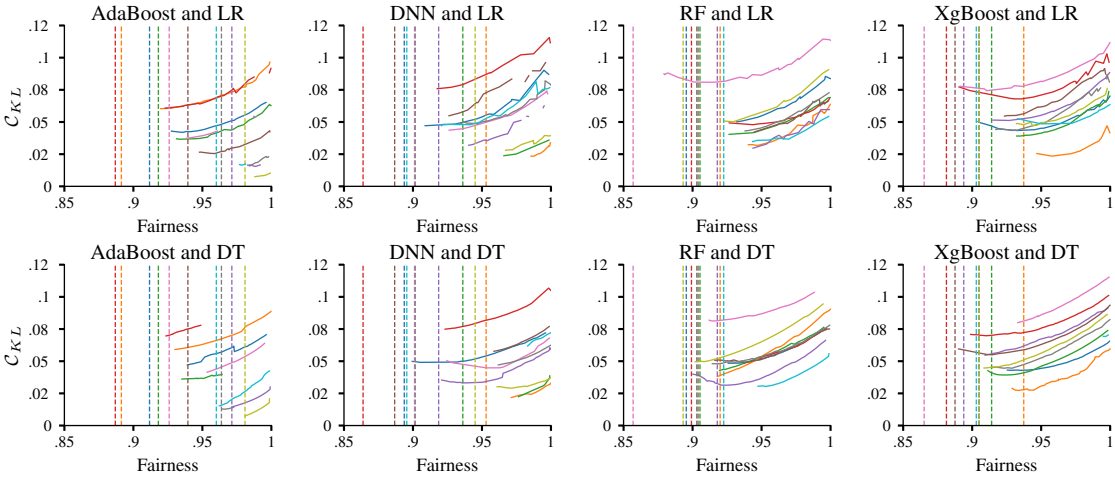

Figure 10: Kullback–Leibler divergence, $\mathcal{C}_{\text{KL}}$, between confusion matrices of subgroups versus demographic parity fairness of two interpretable models, namely Logistic Regression (LR) and Decision Tree (DT), approximating four black-box models, namely AdaBoost, Deep Neural Network (DNN), Random Forest (RF), and Gradient Boosted Decision Trees (XgBoost), using Marketing dataset. We consider 10 split of the dataset to the training set for the black-box model and suing set for the interpretable model. Vertical dashed lines represent unfairness values of the black-box models.

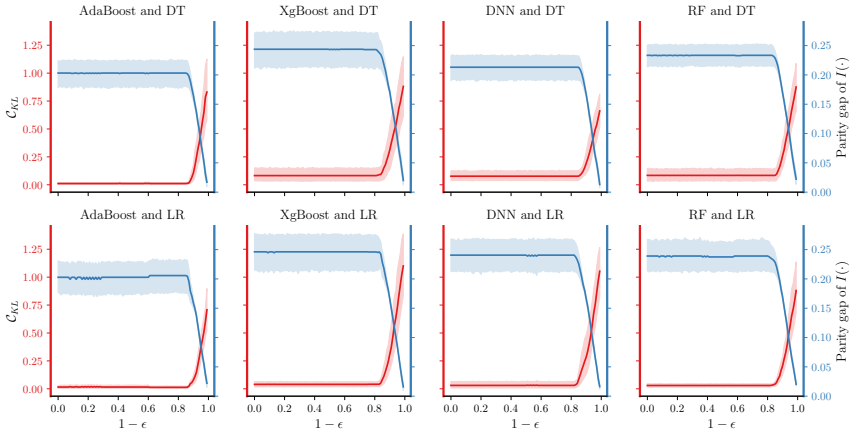

Figure 11: KL among altered confusion matrices of subgroups versus demographic parity unfairness of two explainable models, namely Logistic Regression (LR) and Decision Tree (DT), approximating four black-box models, namely AdaBoost, Deep Neural Network (DNN), Random Forest (RF), and Gradient Boosted Decision Trees (XgBoost), using COMPAS dataset. We consider 10 split of the dataset to the training set for the black-box model and suing set for the interpretable model.