FairCognizer: A Model for Accurate Predictions with Inherent Fairness Evaluation

Adda-Akram Bendoukha, Nesrine Kaaniche, Aymen Boudguiga, Renaud Sirdey

FairCognizer: A Model for Accurate Predictions with Inherent Fairness Evaluation

Adda-Akram Bendoukha [a], **Nesrine Kaaniche** [a], **Aymen Boudguiga** [b] **and Renaud Sirdey** [b]

[a]Samovar, Télécom SudParis, Institut Polytechnique de Paris, France
[b]Université Paris Saclay, CEA List, France

Abstract. Algorithmic fairness is a critical challenge in building trustworthy Machine Learning (ML) models. ML classifiers strive to make predictions that closely match real-world observations (ground truth). However, if the ground truth data itself reflects biases against certain sub-populations, a dilemma arises: prioritize fairness and potentially reduce accuracy, or emphasize accuracy at the expense of fairness. This work proposes a novel training framework that goes beyond achieving high accuracy. Our framework trains a classifier to not only deliver optimal predictions but also to identify potential fairness risks associated with each prediction. To do so, we specify a dual-labeling strategy where the second label contains a per-prediction fairness evaluation, referred to as an unfairness risk evaluation. In addition, we identify a subset of samples as highly vulnerable to group-unfair classifiers. Our experiments demonstrate that our classifiers attain optimal accuracy levels on both the *Adult-Census-Income* and *Compas-Recidivism* datasets. Moreover, they identify unfair predictions with nearly 75% accuracy at the cost of expanding the size of the classifier by a mere 45%.

1 Introduction

Machine learning (ML) systems are becoming increasingly pervasive in our interconnected society, playing a crucial role in various applications, including predictive maintenance, autonomous driving, energy management, and supply chain. Additionally, they extend their influence into judicial, socio-economic, and medical domains, addressing aspects such as loan allocation and recidivism prediction. The broad spectrum of these applications [11, 8, 10, 17] underscores the importance of ensuring fair predictions, especially when historical data contains biases stemming from societal misconceptions [13, 27, 12].

Several examples highlight the prevalence of unfairness and bias, disproportionately impacting minorities, in different critical domains, including eHealth [17, 10] or legal systems [9]. One first example involves skin condition diagnostic tools, which may manifest bias, resulting in misdiagnoses, especially for individuals with darker skin tones [29]. The root of this bias often stems from the under-representation of diverse skin types in the training datasets [32]. This leads to inefficient risk assessments for specific diseases within minority groups, potentially compromising the precision of preventive measures and exacerbating healthcare disparities [10, 8, 17]. In judicial instances, one prominent case revolves around the examination of recidivism risk prediction using the *Propublica* dataset. In [11], Dressel and Farid point out that substantial bias within the predictive algorithms results in significant disparities when applied to black defendants, and shed light on the broader implications of biased AI applications in legal contexts. To address such concerns, regulatory frameworks are being developed. One such effort is the European Union AI act[1]. This act establishes a common legal framework, managing risks based on application type (minimal, limited, high, or unacceptable). It also ensures AI use complies with pre-existing EU regulations, emphasizing fair access and trustworthy AI development.

Several works [9, 22, 14, 15, 36] show that ensuring fairness in supervised learning is often framed as a trade-off between considering a fair representation or an accurate one, in terms of proximity to ground-truth observations. An accurate classifier learns from historical records, generalizing observed statistical patterns to unseen data. However, if these patterns involve many discriminatory records, the classifier will adopt this biased behaviour. Achieving fair training often requires learning an alternative representation of data, generated via pre-processing techniques to remove biases [22, 9, 19, 34]. This alternative representation does not perfectly mirror reality. Consequently, this distributional drift will inevitably degrade the utility of a classifier trained on this alternative fair representation.

In this work, we study the tension between accuracy and fairness from a different perspective. Rather than opting for a trade-off between training a classifier with prevailing accuracy or fairness, we design a dual-objective classifier able to learn both representations (i.e., accurate and fair). It provides an accurate classification associated with the unfairness risk of each prediction it delivers.

Contributions – The contribution of this paper is threefold:

- We introduce FairCognizer, a novel training framework that makes a model learn a dual label $(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}})$; where $\mathcal{Y}_{\text{bin}}$ is the default class label, and $\mathcal{Y}_{\text{fair}}$ is the fair class label. $\mathcal{Y}_{\text{fair}}$ is obtained from a partial de-correlation of $\mathcal{Y}_{\text{bin}}$ from the sensitive attribute $\mathcal{S}$.
- We introduce a novel sample-level unfairness risk measure, characterizing a *Vulnerable-Subset* of records that are highly subject to unfair classifiers. Additionally, we train discriminator classifiers to identify this subset, showing that *vulnerable* records have distinct statistical patterns, enabling them to be accurately classified.
- We conduct experiments on the *Adult-Census-Income* and the *Compas-recidivism-risk* datasets (denoted as Adult and Compas, respectively) that displayed the ability to maintain an optimal accuracy (86% and 68% respectively) while delivering reliable fairness insights, at the expense of increasing the size of the model by

[1] https://artificialintelligenceact.eu

45%.

Paper organization – Section 2 introduces multi-output learning and defines data and classifier (un)fairness. Section 3 reviews the related work and Section 4 describes the proposed dual-label learning framework and presents the main predictive performance metrics. Section 5 introduces a novel sample-level fairness definition. It also extends the proposed framework to identify a subgroup of data records highly vulnerable to unfairness and propose another dual-prediction classifier with optimal accuracy, and a predictive fairness-risk assessment, before concluding in Section 6.

2 Background

In this section, we discuss supervised learning components. Then, we introduce the concept of group fairness in ML and its metrics.

2.1 Supervised Learning

Supervised learning is a fundamental paradigm in machine learning where the algorithm learns from a labeled dataset. The latter consists of paired inputs (features or attributes) and corresponding outputs (labels). By analyzing this dataset, the algorithm learns a mapping function that can generalize to unseen data. In a typical supervised learning scenario, we have a collection of tuples $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where x_i represents the input feature and y_i denotes the output label. The goal is to learn a *hypothesis* function h, such that $h(x_i)$ is as close as possible to y_i for all $x_i \in \mathcal{D}$.

Supervised learning is expressed as an optimization problem, seeking to minimize a loss (or cost) function that quantifies the discrepancy between the prediction and the actual label. The hypothesis function is parameterized, and the learning process involves iterative adjustment of these parameters θ to minimize the loss:

2.2 Multi-output learning

Unlike traditional single-label classification, multi-output classification predicts multiple outputs simultaneously from the same input features. The learning process requires data samples to be labeled accordingly. That is, $\mathcal{D} = \{(x_1, y_1^{(1)}, \ldots, y_1^{(k)}), \ldots, (x_n, y_n^{(1)}, \ldots, y_n^{(k)})\}$ and the optimization is performed on the loss of every output and the corresponding label, as such, at each learning iteration t :

$$L_j(\theta) = \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \mathcal{L}(x_i, y_i^{(j)}, \theta_t) \ \forall j \in \{1, \cdots, k\}$$

$$\text{and} \quad \theta^{t+1} = \mathsf{argmin}_\theta \ L(\theta) = \frac{1}{k} \sum_{j=1}^{k} L_j(\theta^t)$$

For convex loss functions, a common approach for optimization is to use Stochastic Gradient Descent (SGD). This iterative algorithm processes data in batches ($\mathcal{B} \subset \mathcal{D}$) and updates the model parameters (θ) through gradient descent. The learning rate (η) controls the step size of these updates, guiding the parameters towards a minimum of the loss function, such that:

$$g_t = \nabla L(\theta_t) \ \text{ (Gradients computation)}$$
$$\theta_{t+1} = \theta_t - \eta g_t \ \text{ (Parameters update)}$$

2.3 Fairness in Machine Learning

Group fairness investigates the behaviour of an algorithm with respect to inputs belonging to different populations, defined by a sensitive attribute $\mathcal{S}$ that can lead to discrimination, *e.g.*, race or gender. It refers to the statistical independence of the model's prediction $\hat{\mathcal{Y}}$ from the sensitive attribute $\mathcal{S}$.

Removing the sensitive attribute from the training data does not efficiently solve the problem, since indirect discrimination can still occur due to other attributes that act as statistical *proxies* for the sensitive attribute [4, 7]. These proxies can reproduce the same discriminatory behaviour as the sensitive attribute. For example, the zip code (a non-sensitive attribute) is highly correlated with ethnicity (a sensitive attribute) in the USA [7]. Meanwhile, removing all attributes that are correlated to the sensitive one will result in a significant degradation in the model's utility. In addition, artificially removing statistical correlations between sensitive and non-sensitive attributes is challenging, and often results in a fairness-accuracy trade-off [24, 31, 11]. Finally, a noteworthy cause is the unbalance of training datasets w.r.t. underrepresented groups that may lead to poor performance of the model over inputs belonging to these minorities.

In the following, we present main fairness metrics at both data and classifier levels, with respect to a sensitive attribute $\mathcal{S}$[2].

2.3.1 Data unfairness

We review the main metrics [**?**] for evaluating the discrimination in a dataset when considering the joint distribution $(\mathcal{X}, \mathcal{S}, \mathcal{Y})$, such that $\mathcal{X}$ is the set of non-sensitive attributes, and $\mathcal{Y}$ is the class label.

- **Disparate Treatment** refers to the distribution $P(\mathcal{Y}|\mathcal{S})$, and therefore expresses the statistical correlations between the sensitive attribute $\mathcal{S}$ and the label $\mathcal{Y}$ within a dataset. It is often measured as the ratio between the proportion of positively labeled elements from group $\mathcal{S} = s_0$ and group $\mathcal{S} = s_1$:

  $$DT(\mathcal{D}, \mathcal{S}) = \frac{P(\mathcal{Y} = 1|\mathcal{S} = s_0)}{P(\mathcal{Y} = 1|\mathcal{S} = s_1)}.$$

  Therefore, if $\mathcal{S} \perp \mathcal{Y}$ than $DT(\mathcal{D}, \mathcal{S}) = 1$.
- **Disparate Impact** extends the source of discrimination to non-sensitive attributes $\mathcal{X}$, considering them as statistical proxies to the sensitive ones. Hence, it models the distribution $P(\mathcal{S}|\mathcal{X})$. A common measure of disparate impact is quantified from the *Balanced Error Rate* (BER) [14] of the *best* performing adversarial classifier $\bar{f} : \mathcal{X} \to \mathcal{S}$ that infers the attribute $\mathcal{S}$ given $\mathcal{X}$, as:

  $$BER(\bar{f}, \mathcal{S}) = \frac{P(\bar{f}(\mathcal{X}) = 1|\mathcal{S} = s_0) + P(\bar{f}(\mathcal{X}) = 0|\mathcal{S} = s_1)}{2}.$$

  A low $BER(\bar{f}, \mathcal{S})$ indicates a high correlation between $\mathcal{S}$ and $\mathcal{X}$, and therefore, high disparate impact in $\mathcal{D}$.

2.3.2 Classifier unfairness

In this section, we review the main group fairness metrics for evaluating the impact of data unfairness on a classifier's behaviour.

- **Statistical Parity Difference (SPD)** [9] evaluates the proportion of positive outcomes across different groups identified by a sensitive attribute. For example, if males and females have equal qualifications for a job, the proportion of males and females being hired

[2] Without loss of generality, we consider that $\mathcal{S}$ is a binary attribute with values $\{s_0, s_1\}$.

should be roughly the same, i.e.,: $P(\hat{\mathcal{Y}} = 1|\mathcal{S} = s_0) = P(\hat{\mathcal{Y}} = 1|\mathcal{S} = s_1)$. In this case, SPD will be equal to 0 as it satisfies:

$$SPD = |P(\hat{\mathcal{Y}} = 1|\mathcal{S} = s_0) - P(\hat{\mathcal{Y}} = 1|\mathcal{S} = s_1)|$$

However, SPD does not take into account the model's accuracy. In particular, a dummy model that systematically outputs 1 ($h(x) = 1, \forall x$) would perfectly satisfy the statistical parity evaluation.

- **Equal Opportunity Difference (EOD)** [18] quantifies the disparity in true positive rates between groups identified by a sensitive attribute. It compares the odds of receiving a positive outcome for individuals from different groups.

$$EOD = |P(\hat{\mathcal{Y}} = 1|\mathcal{S} = s_0, Y = 1) - P(\hat{\mathcal{Y}} = 1|\mathcal{S} = s_1, \mathcal{Y} = 1)|$$

Other metrics focus on equalizing a model's misperformances across different groups. Hence, measuring disparities in false prediction rates between privileged and protected groups, such as false positive disparity FPD and the false negative one FND.

3 Related work

3.1 Fairness-aware Machine Learning

Three main approaches can be applied to improve group fairness:

Pre-processing These techniques identify and remove bias within a collection of data records before training the models. Kariman and Calders [22] propose two pre-processing strategies: (1) *Massaging the dataset* changes the labels of a subset of data records to remove the bias, and (2) *Reweighing* assigns weights to data records according to their level of fairness with respect to a sensitive attribute. He *et al.* [19] remove Pearson's correlations between every non-sensitive attribute and the set of sensitive ones (disparate impact) by interpreting independence as orthogonality in a vector space of attributes. Maggio *et al.* [25] propose a framework that bridges the gap between the statistical and geometric perspectives on data fairness. They visualize how the distributions of a sensitive feature and the model's predictions are related across different groups. Xu *et al.* [34] use a Generative Adversarial Network (GAN) to train a generator network under two constraints: imitate the true data distribution and remove bias with respect to a sensitive attribute.

In processing Kariman and Calders [22] introduce a sampling approach to make every batch of data that is fed to the classifier in the feedforward phase of the learning (almost) bias-free. Several works [6, 36, 23, 1, 2] add a fairness-related term $\mathcal{F}(\theta, \mathcal{B})$ to the loss function $\mathcal{L}(\theta, \mathcal{B})$ to minimize. Beutel *et al.* [6] define the term $\mathcal{F}(\theta, \cdot)$ as a measure of correlation (mutual information) between the predictions of the model and the sensitive attributes of the data batch $\mathcal{B}$. Zafar *et al.* [36] express the unfairness as an additional optimization constraint: $\mathsf{minimize}\ \ L(\theta, \mathcal{B})\ \mathsf{such\ that}\ : \Omega(\theta, \mathcal{B}) < 0$ where $\Omega(\theta, \mathcal{B}) < 0$ is a fairness-related constraint, ideally expressed as a convex function to maintain the efficiency of the learning process.

Post-processing Works on achieving group fairness after training a classifier aimed at modifying the decision boundaries for different subgroups [5]. For instance, statistical parity between black and white defendants in the Compas dataset can be achieved by carefully lowering the decision threshold of the classifier for black defendants when compared with the threshold for white defendants.

3.2 Accuracy vs fairness trade-off

The impact of fairness-aware training on the predictive performances of a classifier has been widely studied. Most existing research [22, 24, 15, 34, 31] views the trade-off between accuracy and fairness as an unavoidable challenge. However, Wick *et al.* [33] argue that under specific conditions, achieving both high accuracy and fairness might be possible. Kamiran and Calders [22] present a theoretical analysis of this trade-off when learning a fairness-constrained classifier. They establish a linear deterioration of accuracy as fairness improves with pre-processing techniques. Meanwhile, Liu *et al.* [24] handle fairness by formulating it as a multi-objective optimization problem such that the corresponding Pareto fronts express the accuracy-fairness trade-offs. Fish *et al.* [15] propose a generic method called *shifted decision boundary*. This method leverages confidence-based predictions from machine learning models to improve the fairness-accuracy trade-off. It achieves this by strategically adjusting the decision boundary for different groups within the data. Wang *et al.* [31] introduce several fairness measures that capture the multi-task fairness-accuracy trade-offs, and insights on group fairness in multi-task learning settings. In conclusion, most of the proposed solutions show that prioritizing fairness will decrease the model's accuracy.

3.3 Delivering fairness insights along with prediction

Several recent works, including those in individual fairness [35, 16] and explainable machine learning (XAI) [21, 3] have explored the concept of combining fairness information with optimal accuracy predictions. In the context of individual fairness, this concept is often framed as the maximum distance between an input sample x^* and other samples that a classifier f predicts consistently, such that:

$$\max_{\epsilon}\ \ \forall x : d(x^*, x) \leq \epsilon\ \ \text{and}\ \ f(x) = f(x^*)$$

For instance, Yadav *et al.* [35] introduce the concept of a fairness certificate per input. This certificate leverages a distance metric (d) on the most important features of the prediction task. In explainability, measuring fairness often involves the measure of features' impact on the prediction. For example, Jain *et al.* [21] compute *Shapley* values of the sensitive attributes to assess fairness[3].

Maughan *et al.* [26] define the *prediction-sensitivity* measure as the influence of the sensitive attribute on the prediction, and use it to evaluate per-input group-fairness risks. This information is not directly learned from the classifier itself. It is rather computed post-prediction by differentiating the classifier's outputs with respect to the values of the sensitive attributes for a given input x, hence defining a gradients' vector whose norm is proportional to the unfairness level of the classifier at a given input.

In this work, the main challenge is to develop classifiers that are aware of their inherent fairness risk for each input. This risk requires a clear definition and incorporation into the training data while ensuring that classifiers will still learn properly.

4 Towards a dual label fair learning

In this section, we describe our proposed framework for dual-output learning. We assume that we have a single binary sensitive attribute $\mathcal{S} \in \{s_0, s_1\}$ and a set of non-sensitive attributes denoted $\mathcal{X}$ (as presented in Figure 1).

[3] *Shapley* values explain the influence of each feature on the prediction, including those that are sensitive.

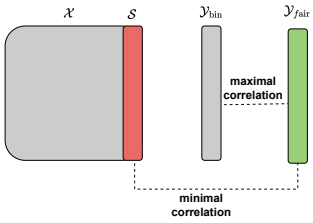


Figure 1: Correlation properties of the generated fairness class labels.

We consider a first binary classification task with $\mathcal{Y} = \mathcal{Y}_{\text{bin}} \in \{0, 1\}$. $\mathcal{Y}_{\text{bin}}$ corresponds to the default classification task of the trained model. Then, we propose to extract a second fairness class label $\mathcal{Y}_{\text{fair}}$, that represents a fair-aware assignation of outcomes with respect to groups $\mathcal{S} = s_0$ or $\mathcal{S} = s_1$.

4.1 Generating $\mathcal{Y}_{fair}$

We formulate the problem of finding fair-aware class labels, as finding an optimal vector $\mathcal{Y}_{\text{fair}}$ that maximizes the correlation with the default labels $\mathcal{Y}_{\text{bin}}$, and minimizes the correlation with the sensitive attribute vector $\mathcal{S}$ (as presented in Figure 1).

We select the Pearson correlation coefficient for generating $\mathcal{Y}_{\text{fair}}$ because it captures the linear relationship between variables. Statistically independent variables have a Pearson correlation of zero, allowing us to model the influence of $\mathcal{S}$ on $\mathcal{Y}_{\text{bin}}$ without incorporating an inherent correlation. Additionally, the linearity property of the Pearson correlation coefficient makes it well-suited for optimization tasks [20]. We recall that Pearson's correlation satisfies:

$$\mathsf{Corr}(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

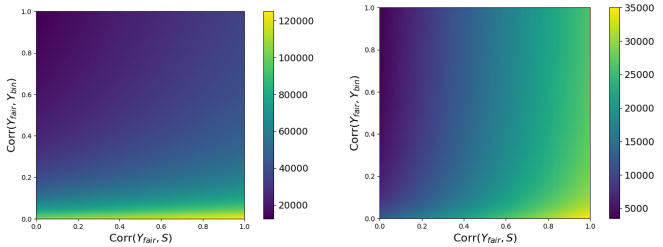
Where $\mathsf{Cov}(X, Y)$ is the covariance of X and Y, and σ_X, σ_Y are respectively standard deviations of X and Y. Correlation values range from -1 to 1. A correlation of both 1 and -1 indicates a deterministic functional relationship between the two variables $Y = f(X)$. A positive correlation indicates that X and Y follow the same variation (increasing f, when the correlation is equal to 1). Conversely, a negative correlation indicates opposite variations of X and Y (decreasing f when the correlation is equal to -1).

We leverage the transitive property of Pearson's correlation[4]. Since $\mathcal{Y}_{\text{bin}}$ reflects the non-sensitive labels ($\mathcal{X}$), maximizing its correlation with the fair prediction ($\mathcal{Y}_{\text{fair}}$) helps maintain the relationship between $\mathcal{Y}_{\text{fair}}$ and $\mathcal{X}$. This leads to a high-performing classifier trained on data containing sensitive attributes ($\mathcal{S}$), non-sensitive attributes ($\mathcal{X}$), and fair predictions ($\mathcal{Y}_{\text{fair}}$). Importantly, minimizing the correlation between $\mathcal{Y}_{\text{fair}}$ and the sensitive attribute ($\mathcal{S}$) reduces bias in the fair predictions. This results in data with less disparate treatment based on the sensitive attribute compared to the original data. These requirements are expressed using the function F_λ defined as:

$$F_\lambda(\mathcal{Y}_{\text{fair}}) = \mathsf{dim}(\mathcal{Y}_{\text{bin}}) \mid \mathsf{Corr}(\mathcal{Y}_{\text{fair}}, \mathcal{S}) \mid + \frac{\lambda}{\mid \mathsf{Corr}(\mathcal{Y}_{\text{fair}}, \mathcal{Y}_{\text{bin}}) \mid} \quad (1)$$

Where λ is a trade-off parameter that reflects the relation between both terms of Equation 1. Figures 2a and 2b show F_λ with $\lambda = 150$ and $\lambda = 300$, respectively. These two values of lambda induce different variations of the function. For example, a variation in the x-axis (i.e., in the correlation of $\mathcal{S}$ and $\mathcal{Y}_{fair}$) has a larger impact on F_λ when $\lambda = 300$.

[4] $\mathcal{Y}_{\text{bin}}$ acts as a statistical proxy of the non-sensitive labels $\mathcal{X}$.



(a) $\lambda = 150$ (b) $\lambda = 300$

Figure 2: Impact of λ on F_λ as a function of $\mid \mathsf{Corr}(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}}) \mid$, and $\mid \mathsf{Corr}(\mathcal{S}, \mathcal{Y}_{\text{fair}}) \mid$ with $\lambda \in \{150, 300\}$

We deduce from the previous observation that higher values of λ result in highly fair-aware class label $\mathcal{Y}_{\text{fair}}$. But, they sacrifice the accuracy of the induced classifier (trained on the $(\mathcal{X}, \mathcal{S}, \mathcal{Y}_{\text{fair}})$ records) due to the loss of useful correlations. Conversely, lower values of λ prioritize the first term and, therefore, produce a highly accurate classifier with slightly improved fairness.

4.2 Optimization strategy

The F_λ function can be categorized as a pseudo-Boolean function according to the PBO definition. A pseudo-Boolean function f is a function that maps a set of binary variables (0 or 1) to real numbers such as $f : \{0, 1\}^n \rightarrow \mathbb{R}$. Several approaches to solving non-linear PBO problems are investigated, including the use of constrained integer programming methods. These methods aim to minimize an objective function subject to constraints on the function variables.

In our case, we introduce hard constraints[5] to limit the search space to the binary space of dimension $|\mathcal{D}|$. We use the Constraint Optimization BY Linear Approximation (COBYLA) solver [28] which is particularly suited for non-linear cost functions with hard constraints. Since this solver does not handle equality constraints, we define the constraints of the boolean solution as two inequality constraints satisfying:

$$\begin{aligned} &\mathsf{minimize}: \ F_\lambda(\mathcal{Y}_{\text{fair}}) \\ &\mathsf{subject\ to}: \ \mathcal{Y}_{\text{fair}}[i] \geq 1 - \epsilon \ \text{or} \ \mathcal{Y}_{\text{fair}}[i] \leq \epsilon \ (\forall i \in [\mathsf{dim}(\mathcal{Y}_{\text{fair}})]) \end{aligned}$$

where $\epsilon = 10^{-5}$ characterizes the constraints on solutions within narrow intervals around the values of 0 or 1. Finally, the solver is run with a maximum of $10k$ iterations with $\mathcal{Y}_{\text{bin}}$ given as the initial guess.

4.3 Learning the dual label

Once the fair class label $\mathcal{Y}_{\text{fair}}$ are generated, the learning objective becomes the mapping: $\mathcal{X}, \mathcal{S} \rightarrow (\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}})$. Indeed, the classifier makes two predictions for each data point (x):

- $\hat{y}_{\text{bin}}$: this prediction focuses solely on accurately matching the default label for the given input (x).
- $\hat{y}_{\text{fair}}$: this prediction represents a fairer and more ethical statistical outcome with respect to the sensitive attribute ($\mathcal{S}$).

This essentially converts the original binary classification task into a multi-output classification. Figure 3 depicts a neural network architecture based on our proposed framework for dual labeling. Applied to the Adultdataset, the figure illustrates how the network achieves

[5] Hard constraints must be satisfied by all variables, while soft constraints only impose penalties on variables that fail to meet them.

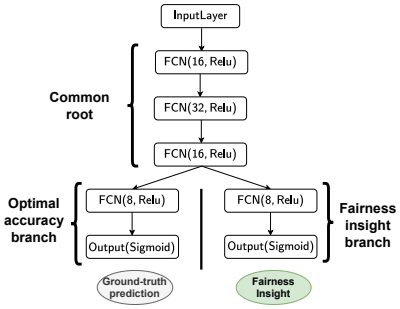


Figure 3: FairCognizer architecture for Adult.

this task with two separate branches. Ideally, both predictions, $\hat{y}_{\text{bin}}$ and $\hat{y}_{\text{fair}}$, would be the same. However, in cases where these predictions differ, the classifier identifies these discrepancies and acts as a per-prediction unfairness alert system, prompting further human evaluation of the specific data point.

4.4 *FairCognizer implementation*

To evaluate the performance of our framework, we conduct various experiments on two datasets: Adult (25k samples) and Compas (5k samples). For our analysis, we focus on the binary groups male and female within these datasets.

4.4.1 *Training analysis*

First, we compute the $\mathcal{Y}_{\text{fair}}$ vector using the COBYLA solver implementation of the SciPy package [30]. We generate $\mathcal{Y}_{\text{fair}}$ for our training subsets, i.e., $\frac{3}{4}$ of both datasets, for the binary groups.

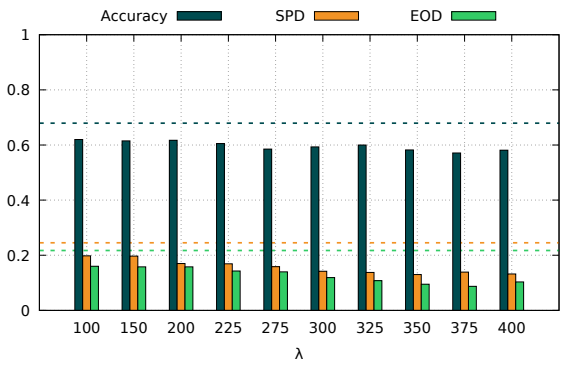
Second, for validation, we train a three-layer classifier on the records $(\mathcal{X}, \mathcal{S}, \mathcal{Y}_{\text{fair}})$ for 10 values of λ, and compare their predictive capabilities and fairness metrics (EOD and SPD) with a baseline classifier trained on the original data. Figure 4 depicts the obtained results. It shows greater fairness improvement using $\mathcal{Y}_{\text{fair}}$ on the Adult dataset compared to Compas. Indeed, our method mainly removes the disparate treatment (direct discrimination), but does not act on the disparate impact (indirect correlation) contained in the dataset. The predominant source of unfairness in the Adult dataset is disparate treatment, whereas in Compas, the primary origin of unfairness is disparate impact[6].

Finally, we train larger FairCognizer classifiers (from 2500 to nearly 3800 trainable parameters) on data-records $(\mathcal{X}, \mathcal{S}, (\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}}))$. We measure their predictive performances and fairness with respect to the initial labels, the fair ones, and the dual-label. Table 1 presents the obtained results on both outputs of the FairCognizer classifier. We note from Table 1 that FairCognizer achieves optimal accuracy for the default class label. This means it prioritizes fairness for the second prediction, $\hat{y}_{fair}$, without compromising accuracy on the original prediction.

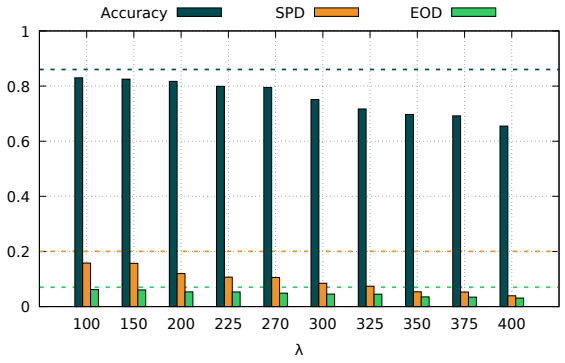
4.4.2 *Classification interpretation*

The four possible predictions made by the dual-label classifier on test data are examined, especially data-records for which the classifier's prediction is $(1, 0)$ or $(0, 1)$ ($\hat{y}_{\text{bin}} \neq \hat{y}_{\text{fair}}$). We measure the prediction inconsistency rate (PIR). That is, we compute the probability: $P(\lceil h(x) \rfloor \neq \lceil h(\overline{x}^{\mathcal{S}}) \rfloor)$, where h is a classifier trained on the

[6] For reference, the BER in Adult for gender as the sensitive attribute $\mathcal{S}$ is 0.183, compared to 0.087 in Compas.



(a) Compas



(b) Adult

Figure 4: Accuracy and fairness measures of classifiers trained on $(\mathcal{X}, \mathcal{S}) \rightarrow \mathcal{Y}_{\text{fair}}$, with $\mathcal{Y}_{\text{fair}}$ generated using different values of λ. Dashed lines represent the baseline measures (color-wise) of a classifier trained on the original dataset.

Table 1: Dual-output model predictive performances with respect to $\mathcal{Y}_{\text{bin}}$, $\mathcal{Y}_{\text{fair}}$ and the dual label $(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}})$, and fairness. $\mathcal{Y}_{\text{fair}}$ is obtained with $\lambda = 250$. Dual-label metrics are average-weighted by the number of true instances in each class.

Measures		Label		
		$\mathcal{Y}_{\text{bin}}$	$\mathcal{Y}_{\text{fair}}$	$(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}})$
Adult	Acc	**0.8543**	0.8258	0.9100
	Precision	0.7308	0.7660	0.8211
	Recall	0.6790	0.6331	0.8182
	f1-score	0.7039	0.6932	0.8196
	SPD	0.1624	**0.1146**	–
	EOD	0.0599	**0.0230**	–
Compas	Acc	**0.6898**	0.6749	0.8388
	Precision	0.6768	0.6823	0.6795
	Recall	0.6098	0.5896	0.6723
	f1-score	0.6415	0.6326	0.6758
	SPD	0.1830	**0.1150**	–
	EOD	0.1510	**0.0879**	–

samples $(\mathcal{X}, \mathcal{S} \rightarrow \mathcal{Y}_{\text{bin}})$ and $\overline{x}^{\mathcal{S}}$ is the sample x where the binary value of $\mathcal{S}$ is flipped. We observe that for the subset of data-records with dual predictions $\hat{y}_{\text{bin}} \neq \hat{y}_{\text{fair}}$, the prediction inconsistency rate is significantly higher compared to data-records for which $\hat{y}_{\text{bin}} = \hat{y}_{\text{fair}}$.

Table 2 shows that the dual-label model can identify unfair predictions, even if they are accurate. These predictions occur for data points similar to the ones from the opposite sensitive group ($\mathcal{S}$), but with different labels. By similar, we refer to close data-records with respect to non-sensitive attributes $\mathcal{X}$. Indeed, flipping the sensitive

Table 2: Prediction inconsistency rates across 1000 sampled data-records for which the fair prediction equals the accurate one, and on records for which the fair prediction is different from the accurate one.

	$(\hat{y}_{\text{bin}} \neq \hat{y}_{\text{fair}})$	$(\hat{y}_{\text{bin}} = \hat{y}_{\text{fair}})$
Adult PIR	58.3%	4.1%
Compas PIR	64.7%	19.0%

attribute value for such a point is likely to flip the model prediction (as indicated by the high PIR values on these points). We conclude that there is a subset of data-records with significantly higher vulnerability to classification unfairness.

5 Identifying vulnerable data-records to unfair classifiers

In this section, we introduce a method to define a subset of data-records as highly vulnerable to group-unfair classifiers identified with the label $(\mathcal{Y}_v)$. Here, our learning objective is the mapping: $\mathcal{X}, \mathcal{S} \rightarrow (\mathcal{Y}_{\text{bin}}, \mathcal{Y}_v)$. So, we specify a classifier to both deliver an optimally accurate prediction and to recognize if the data-record is within its vulnerability region.

Definition 5.1 (Classifier-centric unfairness vulnerability) *A data-record x is said to be vulnerable to group-unfairness of classifier h, and for binary sensitive attribute $\mathcal{S}$ when:*

$$\lceil h(x) \rfloor \neq \lceil h(\overline{x}^{\mathcal{S}}) \rfloor \tag{2}$$

Definition 5.2 (Data-record vulnerability) *The measure of the vulnerability of a data-record x to group-unfairness with respect to a binary sensitive attribute $\mathcal{S}$ relatively to a dataset $\mathcal{D}$ is:*

$$\mathcal{V}_{\mathcal{D}}(x) = P(\lceil h(x) \rfloor \neq \lceil h(\overline{x}^{\mathcal{S}}) \rfloor) \tag{3}$$

where the probability is taken over the set of classifiers h trained on $\mathcal{D}$ without fairness considerations w.r.t. the sensitive attribute $\mathcal{S}$.

Intuitively, the vulnerable data points are those where the sensitive attribute strongly influences the classification. Indeed, these points would likely be classified differently if their sensitive attribute value $\mathcal{S}$ (i.e., prediction on $\overline{x}^{\mathcal{S}}$) belonged to the other group. This unfairness risk measure considers their statistical environment. That is, the overall group unfairness (disparate impact and treatment) within the collection of data-records. This measure allows to define a subset of data-records within a dataset $\mathcal{D}$, for which the value of the sensitive attribute is very likely to determine their classification by a large proportion of classifiers trained on $\mathcal{D}$ without fairness considerations, therefore capturing all existing biases.

To take into consideration the diversity of the learning process and its stochastic nature, we introduce as a parameter, the proportion of classifiers trained on $\mathcal{D}$ for which the samples are vulnerable, which provides a sample-level unfairness risk relative to dataset $\mathcal{D}$. However, the training process is usually numerically stable [7]. That is, converging classifiers result in very closely behaving classifiers in terms of predictions. Therefore, the value $\mathcal{V}_{\mathcal{D}}(x)$ lies either in a small interval near 1 or in a small interval near 0. Therefore, $\mathcal{V}_{\mathcal{D}}(x)$ can be considered as a binary value (vulnerable or not vulnerable).

[7] Using regularization methods (for inputs and parameters), choosing a numerically stable loss function, learning-rate scheduling, and other methods

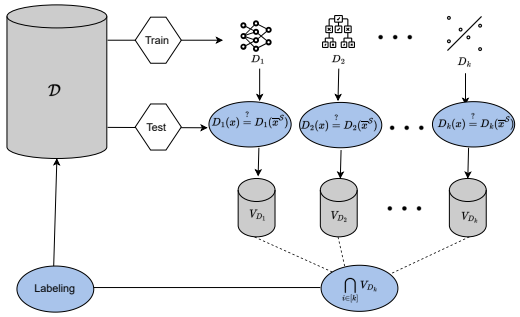


Figure 5: Identification and labeling of data-records in $\mathcal{D}$ as vulnerable or non-vulnerable to group-unfairness with respect to $\mathcal{S}$.

5.1 Vulnerability Labeling

In order to identify the subset of highly vulnerable data-records within $\mathcal{D}$, we train k "discriminator" classifiers $\{D_1, \ldots, D_k\}$ with diverse architectures, suitable for the learning task, and data type, on the prediction task $\mathcal{X}, \mathcal{S} \rightarrow \mathcal{Y}$. The purpose of these classifiers is to form subsets of data-records $V_{D_i} = \{x \in \mathcal{D}$ such that $D_i(x) \neq D_i(\overline{x}^{\mathcal{S}})\}$ associated to discriminator D_i. Afterwards, $\mathcal{V}_{\mathcal{D}} = \bigcap_{i\in[k]} V_{D_i}$ is identified as the set of highly vulnerable data-records.

5.2 Experiments

5.2.1 Vulnerability identification

To identify the subset of vulnerable data-records for sensitive attribute $\mathcal{S} = \{\text{Male}, \text{Female}\}$ in both datasets, we follow the process depicted in figure 5 by training 10 classifiers with diverse architectures: 4 neural networks, with 3, 4, and 5, layers, 3 decision trees with different depths, and 3 SVMs. The first batch of experiments involves the observation, and analysis of the distribution of the vulnerable subsets in both datasets. Subsets of 2351 and 670 samples were identified in Adult and Compas datasets (7% and 9% of the dataset respectively).

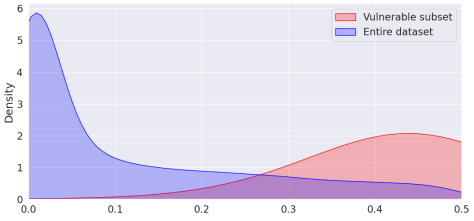
5.2.2 Vulnerable subset analysis

The first analysis of the extracted vulnerable subsets in both Adult, and Compas datasets involved the training of a neural network classifier on these datasets, and the evaluation of its predictive performances and fairness on a baseline test subset from $\mathcal{D}$, on the vulnerable subset $V_{\mathcal{D}}$, and on uniformly sampled non-vulnerable records. Table 3 depicts the obtained results. We observe as well, the classifier's behaviour on these categories of samples through the distribution of the prediction unconfidence $|h(x) - \lceil h(x) \rfloor|$. Furthermore, a substantial increase in the impact of $\mathcal{S}$ on the prediction is observed for samples from $\mathcal{V}_{\mathcal{D}}$ compared to samples from $\mathcal{D} \backslash \mathcal{V}_{\mathcal{D}}$ measured through *Shapley* values.

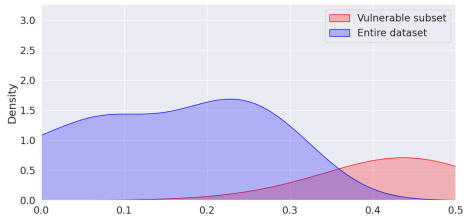
A significant degradation of the classifier's performance, and fairness is observed on the data-records belonging to the vulnerable subset as shown in table 3. Indeed, FND and FPD are significantly higher for samples in $\mathcal{V}_{\mathcal{D}}$, while in $\mathcal{D} \backslash \mathcal{V}_{\mathcal{D}}$, both the predictive performance and the fairness show a notable improvement. This suggests that a large proportion of unfair behaviour with respect to attribute $\mathcal{S}$ takes place in $\mathcal{V}_{\mathcal{D}}$. In addition, the classifier exhibits increased uncertainty in predicting outcomes for vulnerable data records, compared to uniformly sampled records, as shown in Figure 6.

Table 3: Performance, fairness (FPD/FND), and *Shapley* values measures on the vulnerable subset $\mathcal{V}_\mathcal{D}$ compared to a baseline test set from $\mathcal{D}$, and the dataset without the vulnerable subset $\mathcal{D}\backslash\mathcal{V}_\mathcal{D}$.

Measures		Data		
		$\mathcal{D}$	$\mathcal{V}_\mathcal{D}$	$\mathcal{D}\backslash\mathcal{V}_\mathcal{D}$
Adult	Acc	0.8574	0.6293	0.8818
	Precision	0.7528	0.6213	0.8370
	Recall	0.6567	0.6110	0.6616
	FPD	0.1009	0.2771	0.060s1
	FND	0.1860	0.3325	0.0903
	Shapley($\mathcal{S}$)	**0.0233**	**0.0449**	**0.0113**
Compas	Acc	0.6875	0.5582	0.7006
	Precision	0.6760	0.5521	0.7020
	Recall	0.6016	0.8171	0.5770
	FPD	0.2259	0.3512	0.1311
	FND	0.2277	0.3352	0.1459
	Shapley($\mathcal{S}$)	**0.0591**	**0.0981**	**0.0345**



(a) Adult



(b) Compas

Figure 6: Prediction unconfidence density on vulnerable, and uniformly sampled records

Intuition behind the distribution of vulnerable data-records. From a geometric perspective, a classifier that delivers binary predictions, can be seen as a decision boundary in the data space. Equivalently, an $n-1$ dimensional subspace (hyper-plane) that partitions the n-dimensional space [8] (the data space). Data records significantly close, from both sides to the decision boundary are the most vulnerable to group unfairness. In the Adultdataset containing discrimination toward female individuals, a female with an excellent record (*e.g.* very high education level and capital gain) is very unlikely to be discriminated negatively by an unfair classifier ($h(x) = h(\overline{x}^\mathcal{S})$). Conversely, a male with a poor record is also very unlikely to be positively discriminated. Most of the discriminating behaviour of a classifier occurs near the decision boundary. Hence, every classifier defines a vulnerability region. Classifiers trained with fairness considerations are associated with smaller vulnerability regions compared to those trained without any fairness consideration. The distribution of vulnerable data-records can therefore be seen as the subset of strongly average records with respect to important attributes (education level in Adult, number of priors in Compas). The closeness to the average that defines the vulnerability region might represent a group fairness metric of the associated classifier.

[8] Conceptually, this vision corresponds to Support Vector Machines (SVMs), but the geometric perspective applies to other model architectures.

5.2.3 *Learning the vulnerability*

The subset $\mathcal{V}_\mathcal{D}$ is a *"blind spot"* in terms of fairness for classifiers trained on $\mathcal{D}$ without fairness consideration. Indeed, as shown in previous experiments, a large proportion of the unfair behaviour of a classifier (FPD/FND) mainly affects records from this subset. However, since records in $\mathcal{V}_\mathcal{D}$ exhibit statistical patterns that make them highly distinguishable from other samples, it becomes possible to train a classifier that accurately recognizes data-records as vulnerable or not to his own potential unfairness with respect to attribute $\mathcal{S}$. Hence, we follows the same methodology as in 4, and train a dual-label classifier on the mapping $\mathcal{X}, \mathcal{S} \rightarrow (\mathcal{Y}_{\text{bin}}, \mathcal{Y}_v)$. Table 4 depicts the obtained predictive performances.

Table 4: Dual-label models predictive performances with respect to $\mathcal{Y}_{\text{bin}}$, $\mathcal{Y}_v$. and the dual label $(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_v)$. Dual-label metrics are average-weighted by number of true instances in each class.

Measures		Label		
		$\mathcal{Y}_{\text{bin}}$	$\mathcal{Y}_v$	$(\mathcal{Y}_{\text{fair}}, \mathcal{Y}_v)$
Adult	Acc	0.8643	0.9714	0.8796
	Precision	0.7308	0.8855	0.8187
	Recall	0.6828	0.8000	0.8393
	f1-score	0.7059	0.8406	0.8289
Compas	Acc	0.6898	0.9781	0.6727
	Precision	0.6768	0.9751	0.7600
	Recall	0.6098	0.8949	0.7727
	f1-score	0.6415	0.9333	0.7663

Table 4 shows a highly accurate prediction of vulnerability in both datasets, emphasizing the high statistical distinguishability of these data-records from their closeness to the decision boundary. It also shows that predictive performances on the original tasks remain largely intact in dual-label learning, with an accuracy loss of less than 1% compared to optimal levels. Positive second predictions (vulnerability) made by this classifier contain the valuable information that the input data-record is within the *fairness-blind-spot* of this classifier, and therefore, the provided prediction should be carefully scrutinized from a fairness viewpoint.

6 Conclusion

In this work, we explore a novel paradigm of fairness-aware learning that can be succinctly described as follows: If a classifier cannot simultaneously achieve optimal accuracy and group fairness, it can still provide valuable per-prediction insights about fairness risks. We provide two different hybrid pre-processing and in-processing approaches to implement this paradigm. Our study introduces a nuanced analysis of unfairness that encompasses both the classifier and the individual data records. Specifically, individuals exhibit varying degrees of vulnerability to the group-unfairness of a classifier. This nuanced perspective enables a targeted approach for enhancing fairness by directing efforts towards the most susceptible subset of data records affected by classifier unfairness. Our main research perspectives consist of investigating potential privacy risks of FairCognizer, since improving fairness must not harm other aspects of trustworthy machine learning. Indeed, learning extra features from a collection of data-records might result in further leakage, either from the FairCognizer internal weights (white-box leakage), or its dual predictions (black-box leakage).

Acknwoledgement

This work was supported by the France ANR project ANR-22-CE39-0002 EQUIHID.

References

[1] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms, 2019.

[2] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making, 2019.

[3] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E. Gilbert. Can explainable ai explain unfairness? a framework for evaluating explainable ai, 2021.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, MIT Press, 2023.

[5] Joachim Baumann, Anikó Hannák, and Christoph Heitz, 'Enforcing group fairness in algorithmic decision making: Utility maximization under sufficiency', in *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, (June 2022).

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements, 2019.

[7] George J. Borjas, 'Demographic determinants of testing incidence and covid-19 infections in new york city neighborhoods', IZA Discussion Papers 13115, Bonn, (2020).

[8] Irene Chen, Peter Szolovits, and Marzyeh Ghassemi, 'Can ai help reduce disparities in general medical and mental health care?', *AMA journal of ethics*, **21**, E167–179, (02 2019).

[9] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[10] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Duong, and Marzyeh Ghassem, 'Covid-19 image data collection: Prospective predictions are the future', *Machine Learning for Biomedical Imaging*, **1**(December 2020), 1–38, (December 2020).

[11] Julia Dressel and Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances*, **4**(1), eaao5580, (2018).

[12] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto, 'Algorithmic fairness datasets: the story so far', *Data Mining and Knowledge Discovery*, **36**(6), 2074–2152, (September 2022).

[13] Elena Falletti, 'Algorithmic discrimination and privacy protection', *Journal of Digital Technologies and Law*, **1**, 387–420, (06 2023).

[14] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015.

[15] Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. A confidence-based approach for balancing fairness and accuracy, 2016.

[16] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning, 2018.

[17] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits, 'Unfolding physiological state: Mortality modelling in intensive care units', *KDD: Proc Int Con Knowl Discov Data Mining.*, **2014**, 75–84, (08 2014).

[18] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[19] Yuzi He, Keith Burghardt, and Kristina Lerman. Learning fair and interpretable representations via linear orthogonalization, 2019.

[20] Yuzi He, Keith Burghardt, and Kristina Lerman, 'A geometric solution to fair representations', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, p. 279–285, New York, NY, USA, (2020). Association for Computing Machinery.

[21] Aditya Jain, Manish Ravula, and Joydeep Ghosh. Biased models have biased explanations, 12 2020.

[22] Faisal Kamiran and Toon Calders, 'Data pre-processing techniques for classification without discrimination', *Knowledge and Information Systems*, **33**, (10 2011).

[23] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma, 'Fairness-aware learning through regularization approach', in *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650, (2011).

[24] Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach, 2022.

[25] Alessandro Maggio, Luca Giuliani, Roberta Calegari, Michele Lombardi, and Michela Milano, 'A geometric framework for fairness', in *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, October 1st, 2023*, eds., Roberta Calegari, Andrea Aler Tubella, Gabriel González-Castañé, Virginia Dignum, and Michela Milano, volume 3523 of *CEUR Workshop Proceedings*. CEUR-WS.org, (2023).

[26] Krystal Maughan and Joseph P. Near. Towards a measure of individual fairness for deep learning, 2020.

[27] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi, 'The ethics of algorithms: Mapping the debate', *Big Data & Society*, **3**(2), 2053951716679679, (2016).

[28] M. J. D. Powell, 'A direct search optimization method that models the objective and constraint functions by linear interpolation', (1994).

[29] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi, 'Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations', *Nature medicine*, **27**(12), 2176–2182, (2021).

[30] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors, 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', *Nature Methods*, **17**, 261–272, (2020).

[31] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H. Chi, 'Understanding and improving fairness-accuracy trade-offs in multi-task learning', in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, KDD '21. ACM, (August 2021).

[32] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, et al., 'Characteristics of publicly available skin cancer image datasets: a systematic review', *The Lancet Digital Health*, **4**(1), e64–e74, (2022).

[33] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan, 'Unlocking fairness: a trade-off revisited', in *Advances in Neural Information Processing Systems*, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, volume 32. Curran Associates, Inc., (2019).

[34] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks, 2018.

[35] Chhavi Yadav, Amrita Roy Chowdhury, Dan Boneh, and Kamalika Chaudhuri. Fairproof : Confidential and certifiable fairness for neural networks, 2024.

[36] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2017.