Expert Systems With Applications 292 (2025) 128266

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Underrepresentation, label bias, and proxies: Towards Data Bias Profiles for the EU AI act and beyond

Marina Ceccon [a,*], Giandomenico Cornacchia [b], Davide Dalle Pezze [a], Alessandro Fabris [c,d,*], Gian Antonio Susto [a]

[a] Department of Information Engineering, University of Padova, Italy
[b] IBM Research Europe, Dublin, Ireland
[c] Max Planck Institute for Security and Privacy, Bochum, Germany
[d] University of Trieste, Trieste, Italy

ARTICLE INFO

Keywords:
Algorithmic fairness
Anti-discrimination
Bias detection
Data bias
AI act

ABSTRACT

Undesirable biases encoded in the data are key drivers of algorithmic discrimination. Their importance is widely recognized in the algorithmic fairness literature, as well as legislation and standards on anti-discrimination in AI. Despite this recognition, data biases remain understudied, hindering the development of computational best practices for their detection and mitigation.

In this work, we present three common data biases and study their individual and joint effect on algorithmic discrimination across a variety of datasets, models, and fairness measures. We find that underrepresentation of vulnerable populations in training sets is less conducive to discrimination than conventionally affirmed, while combinations of proxies and label bias can be far more critical. Consequently, we develop dedicated mechanisms to detect specific types of bias, and combine them into a preliminary construct we refer to as the *Data Bias Profile (DBP)*. This initial formulation serves as a proof of concept for how different bias signals can be systematically documented. Through a case study with popular fairness datasets, we demonstrate the effectiveness of the DBP in predicting the risk of discriminatory outcomes and the utility of fairness-enhancing interventions. Overall, this article bridges algorithmic fairness research and anti-discrimination policy through a data-centric lens.

1. Introduction

Algorithmic anti-discrimination is a relatively young field, rapidly moving from niche research to market readiness (Álvarez et al., 2024). Several years of work carried out by a growing research community have convincingly shown that algorithms developed without attention to fairness put vulnerable groups at a systematic disadvantage (Angwin, Larson, Mattu, & Kirchner, 2016; Glazko, Mohammed, Kosa, Potluri, & Mankoff, 2024; Obermeyer, Powers, Vogeli, & Mullainathan, 2019). Recognizing the critical implications of this research, policymakers and standards organizations have published regulations and norms on the topic (ISO, 2021; Parliament, 2024; Schwartz et al., 2022). These documents require standardization to apply across many domains where fairness work is critical, including medicine (Obermeyer et al., 2019), finance (Gillis, Meursault, & Ustun, 2024), employment (Fabris et al., 2024), and education (Baker & Hawn, 2022).

Evaluations of data bias are key computational tools for anti-discrimination work. Since biases in the data are fundamental drivers of algorithmic discrimination (Brzezinski et al., 2024; Vetrò, Torchiano, & Mecati, 2021), bias management is mentioned in every recent standard and regulation on algorithmic anti-discrimination (ISO, 2021; Parliament, 2024; Schwartz et al., 2022). Policy formulations on this topic are rather vague, favouring flexibility on one hand, but leaving the contours of law-abiding bias management undefined for practitioners, contributing to legal risk and uncertainty. For example, recent regulation requires providers of AI systems in high-risk domains to signal sufficient efforts of bias detection and mitigation. How this should be done in practice, however, is left completely undefined (Deck, Müller, Braun, Zipperling, & Kühl, 2024).

Defining precise criteria for bias management requires answering several important questions, left mostly unaddressed in the fairness literature. First, different data biases are associated with algorithmic discrimination. Which combinations of biases are more critical for fairness?

* Corresponding authors.
E-mail addresses: marina.ceccon@phd.unipd.it (M. Ceccon), giandomenico.cornacchia1@ibm.com (G. Cornacchia), davide.dallepezze@unipd.it (D. Dalle Pezze), alessandro.fabris@mpi-sp.org, alessandro.fabris@units.it (A. Fabris), gianantonio.susto@unipd.it (G.A. Susto).

https://doi.org/10.1016/j.eswa.2025.128266
Received 19 December 2024; Received in revised form 14 May 2025; Accepted 19 May 2025
Available online 26 May 2025

Second, data biases are typically described qualitatively but vastly lack a quantitative characterization. Is it possible to monitor distinct data biases unambiguously? Third, while a plethora of fairness interventions exist in the literature, a set of guidelines to prioritize them is prominently lacking. In sum, how can practitioners and researchers make decisions about bias mitigation in a principled manner?

Contributions. In this work, we tackle the above questions, providing several contributions.

- We study three types of data bias widely recognized for their negative influence on algorithmic fairness, namely underrepresentation, label bias, and proxies. Through extensive experiments on diverse datasets, algorithms, and fairness measures we analyze their joint influence on algorithmic discrimination. Our experiments show that underrepresentation in training data is overemphasized in the literature while label bias is more critical. This novel result challenges conventional wisdom held in the algorithmic fairness community.
- We propose a principled mechanism for bias detection that is widely applicable in practice. More in detail, we develop a suite of measures to detect specific data biases without auxiliary information from external sources. We integrate these measures into a preliminary construct, the *Data Bias Profile (DBP)*, which provides a quantitative foundation for identifying and communicating data biases, as well as assessing the risk of algorithmic discrimination. This framework serves as a proof of concept, offering a concrete starting point that the research community can build upon to develop a more robust and systematic approach to bias-aware data documentation.
- We discuss the far-reaching implications of our work for researchers and practitioners. We recommend that researchers use DBPs to select datasets with complementary properties for their experiments, overcoming the present limitations of fairness benchmarks. Finally, we make recommendations for practitioners on the curation and utilization of anti-discriminatory datasets.

Structure. Section 2 presents related work. Section 3 introduces data bias and experimental protocols to analyze it. Section 4 studies the effect of data bias. Section 5 describes bias detection, introducing and demonstrating DBPs. Sections 6 and 7 discuss our results in the broader context of algorithmic fairness work concluding with recommendations for researchers and practitioners.

2. Related work

Algorithmic fairness research keeps contributing new approaches to measure the risk of discrimination (Chen, Giudici, Liu, & Raffinetti, 2024; Cooper et al., 2024; Cornacchia et al., 2023; Fabris, Silvello, Susto, & Biega, 2023) and to mitigate it (Cruz & Hardt, 2024; Fajri, Saxena, Pei, & Pechenizkiy, 2024; Yin, Raab, Liu, & Liu, 2023; Zhang, Cheng, Yuan, & Zhang, 2024). Moving fairness toward market readiness requires research on operationalizing algorithmic anti-discrimination policy (Section 2.1), on its close connection with data bias (Section 2.2), and on principled documentation practices to support anti-discrimination (Section 2.3).

2.1. AI policy on anti-discrimination

Influential legislation and standards on anti-discrimination in AI, such as the EU AI Act (Parliament, 2024) and the NIST report on bias in AI (Schwartz et al., 2022) require multi-disciplinary research to translate policy into computational best practices. Drukker et al. (2023), for example, complement NIST guidelines with a list of domain-specific biases that arise in the medical domain. Borgesius, Baranowska, Hacker, and Fabris (2024) focus on the AI Act, summarizing the main requirements for mandatory evaluation of data biases and their documentation. Deck et al. (2024) compile a list of practical challenges for compliance with anti-discrimination requirements outlined in the AI Act. By highlighting the need for model owners to examine "possible biases that are likely to […] lead to discrimination prohibited under Union law", they stress the pressing questions of which biases lead to discriminatory models and what kind of evidence is required to signal sufficient efforts for bias detection and correction. Our work aims to provide flexible computational tools to answer this question across different domains.

2.2. Linking fairness with data properties

A growing line of work centers on quantifying data biases and their influence on models. Guerdan, Coston, Wu, and Holstein (2023) describe five sources of bias affecting target variables and develop a causal framework to disentangle them. Baumann, Castelnovo, Crupi, Inverardi, and Regoli (2023) present several data biases and provide initial insights into their mitigation. Brzezinski et al. (2024) study the variability of fairness measures with respect to the underrepresentation of protected groups and the imbalance between positives and negatives. They postulate certain properties (e.g. "fairness should not vary with underrepresentation") and highlight measures that realize said properties as more "reliable". Fragkathoulas, Papanikou, Karidi, and Pitoura (2024) survey the intersection of fairness and explainability, including explanations that can describe sources of unfairness.

Vetrò et al. (2021) set out to predict the risk of discrimination against vulnerable groups from their underrepresentation in the data. Their work is highly influential for ours; it represents a first attempt at developing mechanisms to detect (a single type of) data bias and connect it with model fairness, opening the way to follow-up studies (Mecati, Torchiano, Vetrò, & De Martin, 2023; Mecati, Vetrò, & Torchiano, 2022). Our manuscript continues this line of work, with important differences in methodology and conclusions. First, we consider three types of data bias, adding label bias and proxies to underrepresentation and studying their joint influence on fairness. Second, we assess model fairness on unbiased test sets. For example, to measure the effect of strong underrepresentation, we remove from the training set a large percentage of items from the disadvantaged group (even 100 %), but we retain them in the test set for a reliable evaluation. Third, we conclude that underrepresentation in training sets is overemphasized in the algorithmic fairness literature and that other data biases can be more critical.

2.3. Quantitative data documentation

Data documentation is increasingly recognized as a central component of trustworthy AI (Fabris, Messina, Silvello, & Susto, 2022; Gebru et al., 2021; Golpayegani et al., 2024; Holland, Hosny, Newman, Joseph, & Chmielinski, 2020; Königstorfer & Thalmann, 2022; Pushkarna, Zaldivar, & Kjartansson, 2022; Rondina, Vetrò, & De Martin, 2023; Sambasivan et al., 2021). With few exceptions, prominent data documentation frameworks are qualitative. Among quantitative frameworks, Holland et al. (2020) emphasize the analysis of correlations between variables to spot anomalous trends. Dominguez-Catena, Paternain, and Galar (2024) develop metrics to quantify representational and stereotypical biases, demonstrating them on a facial expression recognition dataset. In this work, we propose a principled suite of measures to quantify and document biases associated with algorithmic discrimination. We then outline how these can be composed into a preliminary construct, the Data Bias Profile (DBP), to support structured documentation and analysis. Our approach is tailored to one specific aspect of datasets and differs from existing methods, both quantitative and qualitative.

3. Data bias

Data is a fundamental driver of algorithmic discrimination (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2022; Schwartz et al., 2022; Suresh & Guttag, 2021). Data biases are defined as data properties that, if unaddressed, lead to AI systems that perform better or worse for different groups (ISO, 2021). In this section, we describe three types of data

bias widely recognized for their impact on algorithmic fairness along with corresponding bias injection mechanisms.

3.1. Underrepresentation

The term *representativeness* typically refers to the ability of a dataset to support the development of an accurate model for a target population. Underrepresentation of disadvantaged groups in data is described at length in popular media (Cobham, 2020; Perez, 2019) and seminal fairness articles (Buolamwini & Gebru, 2018; Mehrabi et al., 2022; Shankar et al., 2017) as a key driver of algorithmic discrimination. When groups from the target population are underrepresented in training data, it is argued, AI models will fail to generalize and underperform for those groups (Suresh & Guttag, 2021). Indeed, the (under)representation of protected groups in training sets is studied as a predictor of model unfairness (Brzezinski et al., 2024; Vetrò et al., 2021). Influential legislation and standards recognize representation as a central component of algorithmic anti-discrimination (Parliament, 2024; Schwartz et al., 2022) and mandate efforts to document and curb it.

3.2. Label bias

Labels (or target variables) are key to AI. They give machine learning a "ground truth" that models learn to replicate. Since data is a social mirror, labels reflect undesirable disparities in society (Barocas, Hardt, & Narayanan, 2023). Indeed, measurement methods can be biased across protected groups (Vardasbi, de Rijke, Diaz, & Dehghani, 2024). Policing and arrest tend to target poorer neighborhoods, therefore biasing crime data against black US citizens (Bao et al., 2021). Medical data suffers from underdiagnosis due to substandard medical care (Gianfrancesco, Tamang, Yazdany, & Schmajuk, 2018) and barriers to access for vulnerable populations (Obermeyer et al., 2019). Semi-automated labels are especially likely to compound and reinforce spurious biases in training datasets (Jigsaw, 2018). Several methods have been proposed in the literature to counteract unfair label biases under simplifying assumptions (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Kamiran & Calders, 2011; Liu et al., 2024; Wang, Liu, & Levy, 2021). Overall, measurement bias is inevitable (Jacobs & Wallach, 2021); it is especially problematic for anti-discrimination when it tilts target labels against a vulnerable group (Mehrabi et al., 2022; Suresh & Guttag, 2021). Models trained to predict these labels will encode the underlying biases and harm disadvantaged groups (Bao et al., 2021; Obermeyer et al., 2019).

3.3. Proxies

Proxies are features that correlate with protected attributes. Protected attributes such as gender can be revealed by individual features, such as names in a resume (Santamaría & Mihaljevic, 2018), or by combinations of features, such as the browsing history of a person (Hu, Zeng, Li, Niu, & Chen, 2007). Pursuing fairness by simply removing protected attributes from input features, an approach termed *fairness through unawareness*, is ineffective precisely for this reason: a redundant encoding of latent protected variables is present in other features (Barocas & Selbst, 2016; Hardt, Price, & Srebro, 2016; Pedreschi, Ruggieri, & Turini, 2008). Proxy removal, for example through feature selection or projection, is a popular approach to improve algorithmic fairness (Alves, Amblard, Bernier, Couceiro, & Napoli, 2021; Blind Stairs, 2024; Edwards & Storkey, 2016; HireVue, 2022; Madras, Creager, Pitassi, & Zemel, 2018). Conversely, input features that are strongly correlated with protected attributes are considered a driver of unfairness in data-driven models (Schwartz et al., 2022). Policymakers may therefore expect practitioners to actively identify and eliminate strong proxy features from models powering automated decisions (Bogen, 2024).

Table 1
Notation. Main notational convention adopted in this work.

symbol	meaning
$s \in \mathcal{S}$	protected attribute
$s = d$	historically disadvantaged group
$s = a$	historically advantaged group
$y \in \mathcal{Y}$	target variables
$\mathcal{Y} = \{\oplus, \ominus\}$	positive and negative target values
$x \in \mathcal{X}$	non-protected attributes
$\hat{y} = g(x)$	estimation of y through classifier $g(\cdot)$
$\sigma = \{(x_i, y_i, s_i)\}$	a training set
$\Pr_\sigma(s = d)$	prevalence of d in set σ
$\sigma_d = \{i \in \sigma \mid s_i = d\}$	subset of σ with all data points from group d
σ', y'	training set and target labels after bias injection
$r \in (0, 1)$	percentage of disadvantaged instances retained for training: $\Pr_{\sigma'}(s = d) = r \cdot \Pr_\sigma(s = d)$
$u = r - 1$	underrepresentation factor
$f \in (0, 1)$	flip factor or label bias: $f = \Pr(y'_i = \ominus \mid y_i = \oplus, s_i = d)$

3.4. Data bias injection

Notation. Table 1 summarizes the notational conventions. We let $s \in \mathcal{S}$ denote a sensitive attribute,[1] with value $s = a$ ($s = d$) denoting the historically (dis)advantaged group. We let y indicate the target variable with values in $\mathcal{Y} = \{\oplus, \ominus\}$ and we let $x \in \mathcal{X}$ denote the non-protected features used for classification. Target variables are estimated through a classifier $\hat{y} = g(x)$. We let $\sigma = \{(x_i, y_i, s_i)\}$ denote a sample and $\Pr_\sigma(s = d)$ indicate the prevalence of items with $s_i = d$ in that sample. To inject biases in training sets, we subsample the disadvantaged group and flip its labels. We use σ' to denote a training set derived from σ via bias injection.

Underrepresentation. We let $r \in (0, 1)$ denote the percentage of instances from the disadvantaged group retained for training, so that

$$\begin{aligned} \Pr_{\sigma'}(s = d) &= r \cdot \Pr_{\sigma}(s = d) \\ u &= r - 1. \end{aligned} \tag{1}$$

We call $u = 1 - r$ the *underrepresentation factor* for the disadvantaged group. We vary u across its full range; extreme values $u = 1$ and $u = 0$ denote complete underrepresentation and no underrepresentation, respectively.

Label bias. For label bias, we selectively flip labels. We let $f \in (0, 1)$ indicate the proportion of positive instances ($y = \oplus$) from the vulnerable group whose label is flipped to negative ($y = \ominus$), i.e.

$$f = \Pr(y'_i = \ominus \mid y_i = \oplus, s_i = d). \tag{2}$$

We let the *flip factor* f vary between $f = 0$ and $f = 1$; the former corresponds to no bias injection, the latter to maximum bias where all the positive items from the disadvantaged group in the training set are flipped to a negative target label.

Proxies. We quantify the strength of proxies as their joint ability to predict sensitive attributes. We train a classifier $\hat{s} = h(x)$ to estimate the protected attribute s and we compute its AUC to measure the strength of proxies.

$$\begin{aligned} \hat{s} &= h(x) \\ s\text{AUC} &= \text{AUC}(h). \end{aligned} \tag{3}$$

We term sAUC the *proxy factor* and propose two mechanisms to vary it. An additive protocol adds to the non-sensitive variables $\mathcal{X}$ a new feature correlated with sensitive variables

$$\begin{aligned} x_{\text{new}} &= s + v, \quad v \sim \mathcal{N}(0, \text{std}^2) \\ \mathcal{X}' &= \mathcal{X} \times \mathcal{X}_{\text{new}}, \end{aligned} \tag{4}$$

[1] We use the nomenclature *sensitive* and *protected* attribute interchangeably.

Table 2
Dataset basics. We report dataset name, sensitive attribute information such as (dis)advantaged groups and their prevalence, and target variables information such as positive classes and their prevalence among members of the advantaged and disadvantaged groups.

Dataset	s	a	d	y	$y = \oplus$	$\mathbf{Pr}_\sigma(s = a)$	$\mathbf{Pr}_\sigma(y = \oplus \mid s = a)$	$\mathbf{Pr}_\sigma(y = \oplus \mid s = d)$
Adult	Gender	Male	Female	income	> 50K	0.68	0.3125	0.1136
Adult	Marital status	Married	Not married/ Divorced	income	> 50K	0.48	0.4451	0.0668
Compas	Ethnicity	Caucasian	African-American	recidivism	no reoffense	0.60	0.6091	0.4769
Crime	Ethnicity	Caucasian	Other	violent crime rate	low	0.58	0.7335	0.1805
Folktables	Ethnicity	Caucasian	Other	employment	employed	0.89	0.5688	0.5048
German	Age	> 25 y	<= 25 y	credit risk	good	0.81	0.7284	0.5789
NIH	Gender	Male	Female	chest pathologies	presence of pathology	0.54	0.4112	0.3981
Fitzpatrick17k	Skin type	Light	Dark	skin conditions	presence of condition	0.86	0.2826	0.1897

where v is a normal random variable with zero mean and std^2 variance; we increase the strength of proxies by reducing std. In addition, we consider a subtractive protocol, iteratively removing from the non-sensitive variables $\mathcal{X}$ the strongest predictors of the sensitive attribute

$$x_{\text{drop}} = \max_x \text{sim}(x, s)$$
$$\mathcal{X}' = \mathcal{X} \setminus \mathcal{X}_{\text{drop}}, \tag{5}$$

where $\text{sim}(\cdot,\cdot)$ denotes a similarity function (e.g. correlation) and $\mathcal{X} \setminus \mathcal{X}_{\text{drop}}$ is the feature space obtained removing x_{drop} from the original feature set.[2]

4. Effect of data bias

In this section, we investigate the effect of data biases. We inject data bias into the training and validation datasets of classification models and assess their combined influence on algorithmic discrimination, evaluating fairness metrics on unbiased test sets.

4.1. Overall setup

Datasets. We consider five tabular and two medical imaging datasets, described in Table 2. These datasets are popular in the fairness literature and span several domains where fairness work is critical. Datasets contain information on protected attributes including gender, age, ethnicity, and marital status. Table 2 additionally reports the target variable of each dataset, the prevalence of the disadvantaged and the advantaged groups, as well as the prevalence of positive items in each group. The advantaged group has a higher rate of positive samples compared to the disadvantaged group. Notice that the positive class indicates more desirable outcomes for the assessed individuals, insofar as it is associated with critical resource allocation (loans, medical attention) or lower penalties (incarceration, strict policing). This makes high true positive rates unambiguously important to counter undesirable patterns harming disadvantaged groups, such as underallocation and overcriminalization. More details on each dataset are reported in Appendix A.

Models. We train deep learning models for medical imaging datasets and traditional machine learning models for tabular data. Models optimize accuracy-oriented loss functions without any fairness-enhancing component. For each of the tabular datasets, we train a random forest (RF), a support vector classifier (SVC), and a linear regression (LR). Following the literature, we train a Densenet121 on NIH and a vgg16 model on Fitzpatrick17k (Groh et al., 2021; Seyyed-Kalantari, Liu, McDermott, Chen, & Ghassemi, 2020).

Splits & repetitions. We process tabular datasets with an 80-10-10 train-validation-test split. For NIH, we follow the literature with an 80-10-10 train-validation-test split (Seyyed-Kalantari et al., 2020). For Fitzpatrick17k, we use a 70-15-15 split to ensure sufficient representation of the disadvantaged group in the test set, favoring more stable fairness measurements. After splitting the data, we inject biases in the training and validation set. We keep the test set *unbiased* for a reliable evaluation.[3] For each experiment, we perform 10 training repetitions (with different initial seeds), reporting the mean and standard deviation for metrics of interest.

Performance metrics. To evaluate the classification performance of each model across (often imbalanced) datasets, we consider the balanced accuracy on the test set, i.e. the average between the true positive rate and the true negative rate:

$$\text{BA}_\sigma = \frac{\Pr_\sigma(\hat{y} = \oplus \mid y = \oplus) + \Pr_\sigma(\hat{y} = \ominus \mid y = \ominus)}{2}. \tag{6}$$

Fairness metrics. To assess the model fairness, we consider three complementary metrics: demographic parity, equal opportunity, and predictive quality parity. Demographic parity (DP), also called independence (Barocas et al., 2023), and instantiated as a mean difference (Zliobaite, 2017), is defined as the difference between the acceptance rates computed on different groups

$$\text{DP}_\sigma = \Pr_\sigma(\hat{y} = \oplus \mid s = a) - \Pr_\sigma(\hat{y} = \oplus \mid s = d) \tag{7}$$

and it is independent of the ground truth labels. It is especially salient in contexts where reliable ground truth information is hard to obtain and a positive outcome is desirable, including employment, credit, and criminal justice (Du, Yang, Zou, & Hu, 2021; Gajane & Pechenizkiy, 2018).

Contrary to DP, the equal opportunity (EO) metric is based on the target variable y (Hardt et al., 2016); it is defined as the difference in the true positive rates:

$$\text{EO}_\sigma = \Pr_\sigma(\hat{y} = \oplus \mid s = a, y = \oplus) - \Pr_\sigma(\hat{y} = \oplus \mid s = d, y = \oplus). \tag{8}$$

EO is especially important in contexts, such as healthcare, where a ground truth of reasonable accuracy is available and false negatives (missed diagnosis) are especially harmful.

A third anti-discrimination criterion, focused on both types of misclassification, is represented by prediction quality parity (PQP) (Du et al., 2021). We define it as the difference in balanced accuracy between sensitive groups:

$$\text{PQP}_\sigma = \text{BA}_{\sigma_a} - \text{BA}_{\sigma_d}. \tag{9}$$

In the experiments below, we measure algorithmic fairness according to these metrics as we inject controlled biases the training sets. We present results for logistic regression (on tabular datasets) and equal opportunity, which are representative of broader trends across all models and metrics. The results for the remaining models and metrics can be found in Appendix B; unless explicitly stated, they are equivalent to those illustrated below.

4.2. Underrepresentation

Setup. To study the effect of underrepresentation, we train models in four different settings, varying the underrepresentation factor by

[2] Eq. (5) is a single iteration of the subtractive protocol.

[3] We use the term *unbiased* in a narrow sense, to denote simple random samples of the original dataset, as opposed to subsets where we deliberately inject different types of data bias.

Table 3
Model fairness is mostly unaffected by underrepresentation in the training set. Equal Opportunity (EO), varying the underrepresentation of the minority group in the training set from $u = 0$ (no bias) to $u = 1$ (maximum bias). Mean and standard deviation over 10 repetitions. Symbols (*) and (**) denote statistically significant differences with respect to $u = 0$ at $p = .05$ and $p = .01$, respectively, measured with an unpaired t-test.

			EO			
Dataset	sensitive	model	$u = 0$ (no bias)	$u = 0.2$	$u = 0.8$	$u = 1$ (max bias)
Adult	gender	LR	0.08 ± 0.02	0.09 ± 0.02	0.09 ± 0.02	0.21 ± 0.07**
Adult	marital-status	LR	0.36 ± 0.03	0.36 ± 0.03	0.37 ± 0.03	0.29 ± 0.04**
Compas	race	LR	0.17 ± 0.03	0.17 ± 0.03	0.16 ± 0.02	0.16 ± 0.02
Crime	race	LR	0.33 ± 0.11	0.36 ± 0.12	0.30 ± 0.14	0.28 ± 0.13
Folktables	race	LR	0.05 ± 0.04	0.05 ± 0.04	0.05 ± 0.04	0.05 ± 0.04
German	age	LR	0.07 ± 0.13	0.06 ± 0.11	0.07 ± 0.09	0.06 ± 0.07
NIH	gender	DenseNet	0.01 ± 0.02	0.04 ± 0.01**	0.03 ± 0.02	0.06 ± 0.02**
Fitzpatrick17k	skin type	vgg16	0.09 ± 0.02	0.11 ± 0.02	0.11 ± 0.02	0.11 ± 0.02

undersampling the disadvantaged group. We consider the unbiased case ($u = 0$), the fully biased case ($u = 1$), where the disadvantaged group is completely absent from the training set, and two intermediate settings ($u = 0.2$, $u = 0.8$).

Results. Remarkably, Table 4 shows that the underrepresentation of the minority group does not have a strong impact on fairness: EO is approximately constant as u varies in all datasets. The only datasets for which the increase in disparity is statistically significant are NIH and Adult (gender). For Crime and Adult (marital status), the gap even decreases slightly when the disadvantaged group is removed from the training set.

This trend is surprising and contradicts popular narratives about the effect of underrepresentation on algorithmic fairness. To further analyze these results, we split EO into its groupwise TPR components (Eq. (8)). Fig. 1 reports boxplots of the TPR for the advantaged and disadvantaged groups in NIH and Folktables, which are representative of the remaining datasets. Fig. 1 shows that the TPR remains approximately stable as underrepresentation u varies maximally. Specifically, the TPR for both the advantaged and disadvantaged groups in Folktables are perfectly stable, while they slightly decrease for NIH. The decrease is more marked for the disadvantaged group, leading to a small increase in EO. Underrepresentation is more impactful for NIH since nearly half of the original training set consists of points from the disadvantaged group (Table 2).

Interpretation. This notable result contradicts the position commonly held in algorithmic fairness that increasing the representation of disadvantaged groups in training sets is critical for equitable outcomes. We defer a broader interpretation of this result to Section 6, where we discuss our findings in the broader context of algorithmic fairness research and practice. For now, we highlight this as an indication that underrepresentation in training sets is overemphasized and that other biases may be stronger drivers of model unfairness.

4.3. Label bias

Setup. In this section, we train models on data affected by different degrees of label bias. Specifically, we take a portion (f) of positive samples from the disadvantaged group in the training set and flip their labels to negative. We let the flip factor (Eq. (2)) take values $f = 0$ (no bias), $f = 0.2$ (moderate bias), $f = 0.8$ (strong bias), and $f = 1$ (maximum bias). Additionally, we study the interplay between label bias and underrepresentation by analyzing a scenario in which the prevalence of the disadvantaged group is decreased and part of its positives are flipped. More in detail, at the end of this section, we assess the joint effect of a weak label bias ($f \in \{0, 0.2\}$) and widely-ranging underrepresentation ($u \in \{0, 1\}$).

We report the mean and standard deviation of EO across ten repetitions. As in the previous section, we focus on LR, while results for other models are available in Appendix B along with additional fairness measures.

Results. Table 4 shows that label bias has a large impact on fairness, sizably stronger than underrepresentation (Table 3). Indeed, across all experiments, unfairness grows as f increases. In datasets like NIH and Adult, this increase is very sizable, while for others such as Folktables it is more contained. Tables B.15, B.18, and B.19 in Appendix C show

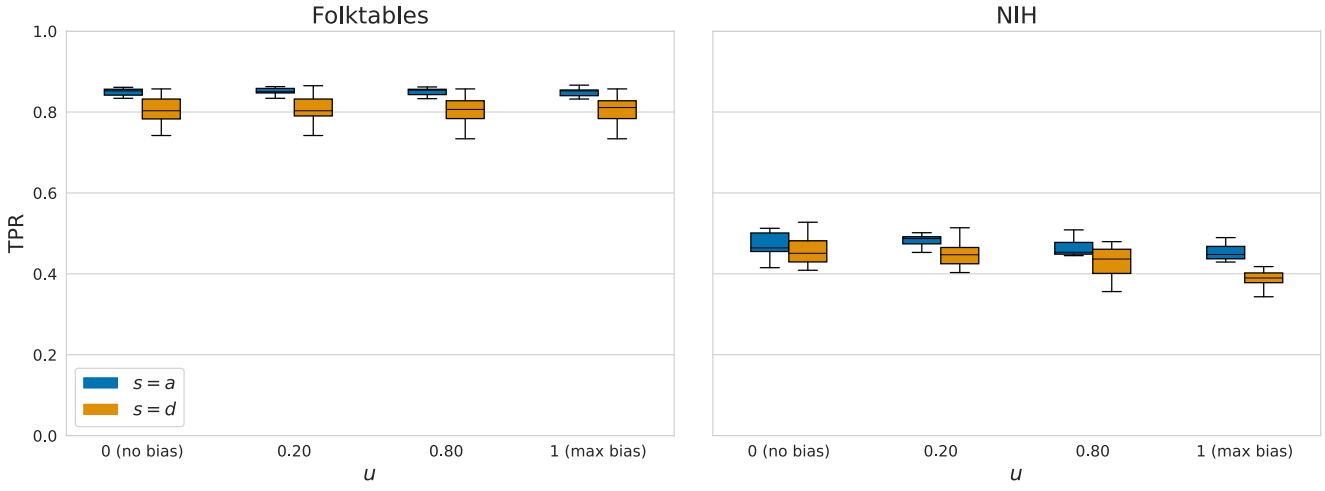

Fig. 1. Large underrepresentation induces minor variations in the True positive rates (TPR) of both groups. Boxplots represent the TPR of the advantaged ($s = a$) and disadvantaged group ($s = d$), as u varies.

Table 4
Model fairness is strongly affected by label bias. Equal Opportunity (EO), as the percentage of flipped positives in the disadvantaged group varies from $f = 0$ (no bias) to $f = 1$ (maximum bias). Mean and standard deviation over 10 repetitions. Symbols (*) and (**) denote statistically significant differences with respect to $f = 0$ at $p = .05$ and $p = .01$, respectively, measured with an unpaired t-test.

			EO			
Dataset	sensitive	model	$f = 0$ (no bias)	$f = 0.2$	$f = 0.8$	$f = 1$ (max bias)
Adult	gender	LR	0.08 ± 0.02	0.21 ± 0.02**	0.52 ± 0.03**	0.55 ± 0.02**
Adult	marital-status	LR	0.36 ± 0.03	0.43 ± 0.04**	0.62 ± 0.02**	0.63 ± 0.02**
Compas	race	LR	0.17 ± 0.03	0.21 ± 0.05	0.21 ± 0.05	0.15 ± 0.05
Crime	race	LR	0.33 ± 0.11	0.39 ± 0.11	0.62 ± 0.17**	0.67 ± 0.15**
Folktables	race	LR	0.05 ± 0.04	0.05 ± 0.04	0.08 ± 0.04	0.09 ± 0.04
German	age	LR	0.07 ± 0.13	0.10 ± 0.14	0.22 ± 0.16	0.25 ± 0.10**
NIH	gender	DenseNet	0.01 ± 0.02	0.09 ± 0.02**	0.40 ± 0.03**	0.48 ± 0.01**
Fitzpatrick17k	skin type	vgg16	0.09 ± 0.05	0.16 ± 0.06*	0.21 ± 0.08**	0.28 ± 0.02**

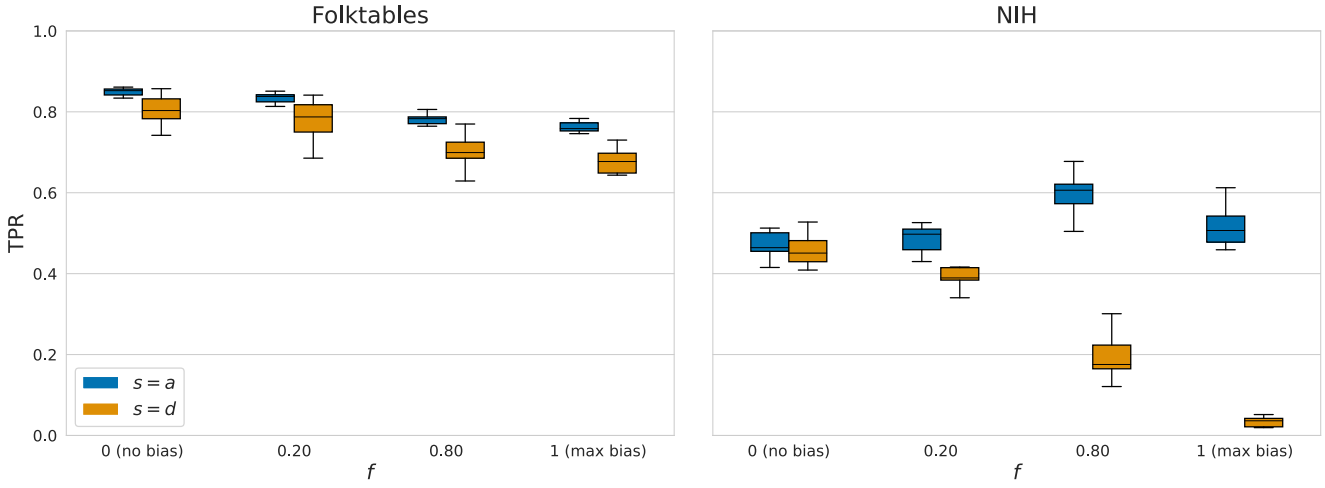

Fig. 2. Label bias induces sizable variations in groupwise true positive rates (TPR); the disadvantaged group is especially affected. Boxplots representing the TPR on the advantaged and disadvantaged group (y axis), as the percentage of disadvataged group items with flipped labels increases in the training set (x axis).

a large increment in unfairness across the remaining metrics (PQP and DP) and models (random forests and SVC).

Zooming in on this result, Fig. 2 depicts the TPR for both sensitive groups on Folktables and NIH. We observe that label bias has a significant impact on the TPR of the disadvantaged group in both datasets, a trend observed consistently across all considered datasets. On the other hand, the impact on the TPR of the advantaged group differs between datasets. Fig. 2 shows a relatively stable TPR for NIH, while, for Folktables, the TPR of the advantaged group decreases with f. Results for the remaining datasets are reported in Tables B.17 and B.16 in the appendix.

Broadly speaking, we distinguish two categories of datasets based on the effect of f on the TPR of the advantaged group. Datasets such as Adult, Crime, Fitzpatrick17k, and NIH exhibit stable values for the TPR of the advantaged group while the TPR of the disadvantaged group decreases, therefore widening the gap. Conversely, datasets like German and Compas show patterns akin to Folktables, resulting in a less pronounced TPR gap. This diverging behavior is explained in Section 4.4.

On the joint effect of label bias and underrepresentation. Table 4 suggests that even a weak label bias can have a sizable impact on model fairness. We therefore study the joint effect of a weak label bias ($f \in \{0, 0.2\}$) and widely-ranging underrepresentation ($u \in \{0, 1\}$). Specifically, Table 5 summarizes the impact of excluding the disadvantaged group from the training set by presenting the difference between fairness under maximum underrepresentation ($\text{EO}_{u=1}$) and no underrepresentation ($\text{EO}_{u=0}$):

$$\Delta\text{EO} = \text{EO}_{u=1} - \text{EO}_{u=0}.$$

Positive values of ΔEO indicate that the inclusion of the disadvantaged group in the training set ($u = 0$) leads to a decrease in the EO metric and, therefore, a relative improvement in their TPR. We quantify this improvement under no label bias ($f = 0$) and weak label bias ($f = 0.2$). As discussed in Section 4.2, underrepresentation of the disadvantaged group in the training set (without label bias) has no clear effect on fairness, as confirmed by the first column of Table 4 ($f = 0$) displaying both positive and negative values. On the other hand, the second column ($f = 0.2$) consistently displays negative values (with the exception of NIH and Compas). This means that, in the presence of relatively weak label bias, it may become preferable for the disadvantaged group to be *completely omitted* from the training set.

Increasing the representation of the disadvantaged group in the training set under these conditions is not only unbeneficial but can, in some cases, be detrimental.

Interpretation. The results presented in this section underscore the critical importance of precise and well-curated ground truth labels in datasets used for training classification models. Specifically, if the labels associated with one demographic group contain noise due to structural discrimination, this can significantly impact model fairness, thereby exacerbating existing biases. Our findings indicate that model performance for the disadvantaged group is consistently and substantially affected when the input labels for this group are subject to systematic bias. Conversely, label bias against the disadvantaged group has a weaker impact on the advantaged group, especially when proxies are strong (see Section 4.4); this discrepancy invariably leads to a fairness decline. Furthermore, we showed that even a small proportion of flipped labels can negatively affect the TPR gap. Overall, this shows that hastily adding disadvantaged groups into training sets without careful label curation can cause more harm than good for the members of those groups.

Table 5
In the presence of weak label noise (f = 0.2), it becomes preferable to omit the disadvantaged group from the training set. The table below illustrates the Equal Opportunity difference (ΔEO) between no representation and full representation for the disadvantaged group across two scenarios: one without label noise ($f = 0$) and one with weak label noise ($f = 0.2$). Negative values indicate a relative decline in the TPR of the disadvantaged group.

			ΔEO	
Dataset	sensitive	model	$f = 0$	$f = 0.2$
Adult	gender	LR	0.13 ± 0.07	−0.01 ± 0.08
Adult	marital-status	LR	−0.07 ± 0.05	−0.13 ± 0.06
Compas	race	LR	0.01 ± 0.01	0.02 ± 0.04
Crime	race	LR	−0.05 ± 0.17	−0.11 ± 0.17
Folktables	race	LR	0.00 ± 0.06	−0.01 ± 0.06
German	age	LR	−0.01 ± 0.15	−0.09 ± 0.17
NIH	gender	DenseNet	0.05 ± 0.03	0.03 ± 0.03
Fitzpatrick17k	skin type	vgg16	0.02 ± 0.07	−0.05 ± 0.08

Table 6
The datasets with strong proxies are most affected by label bias. Strength of proxies for all datasets as measured by *s*AUC across 10 repetitions. We report sample means and standard deviations. The datasets with high *s*AUC values, such as Adult, Crime and NIH (bold), display the highest unfairness under label bias (Table 4).

Dataset	sensitive	model	sAUC
Adult	**gender**	**LR**	**0.9349** ± 0.0025
Adult	**marital-status**	**LR**	**0.9893** ± 0.0011
Compas	race	LR	0.6940 ± 0.0140
Crime	**race**	**LR**	**0.9847** ± 0.0074
Folktables	race	LR	0.6821 ± 0.0112
German	age	LR	0.7939 ± 0.0382
NIH	**gender**	**DenseNet**	**0.9979** ± 0.0013
Fitzpatrick17k	skin type	vgg16	0.8946 ± 0.0100

4.4. Proxies

Setup. In this section, we study the effect of proxies on model fairness. We train classifiers $\hat{s} = h(x)$ to estimate the protected attribute s and we compute their AUC to measure the strength of proxies (*s*AUC – Eq. (3)).[4] To vary the strength of proxies, we leverage the subtractive protocol introduced in Section 3.4 by iteratively removing the feature that is most correlated with the protected attribute; we train a new classifier on the reduced input space and repeat this process until a random classifier performance is reached. We study proxies in conjunction with label bias ($f \in (0, 1)$).

Results. Table 6 reports *s*AUC values for all datasets. Based on these values, we distinguish two types of datasets with diverging properties. Datasets like Adult (gender), Adult (marital-status), Crime, and NIH have very strong proxies for sensitive attributes (*s*AUC > 0.9), while Compas and Folktables have weaker proxies (*s*AUC < 0.7), indicating less information on sensitive attributes encoded in non-sensitive ones. This division mirrors Table 4, where the four high-proxy datasets have the worst (highest) values for EO under maximum label bias ($f = 1$) and three out of four — i.e. Adult (gender), Adult (marital-status), and NIH — already show statistically significant differences under low flip factors ($f = 0.2$). Conversely, the negative effects of label bias are weaker for low-proxy datasets: both Folktables and Compas have no significant differences in EO even under maximum label bias ($f = 1$).

To further investigate this trend, we jointly study label bias and proxies. Fig. 3 depicts EO for Adult (gender) and Folktables as the flip factor f increases. Curves with different colors represent input spaces whose size (number of features n) is iteratively reduced by one. For both datasets, the impact of flips on EO (summarized by average curve slopes) decreases with *s*AUC. The decrease is more marked for Adult since it has stronger proxies and therefore starts from higher *s*AUC values, as reported in the legend.

Interpretation. These results prove that the presence of strong proxies amplifies the risk of algorithmic discrimination, particularly under label bias; when *s*AUC is large, the removal of correlated features can mitigate this risk by reducing the model's reliance on sensitive information. These findings highlight the critical role of proxy strength in exacerbating label bias and influencing the effectiveness of fairness interventions.

5. Bias detection

In this section, we introduce mechanisms for bias detection. We evaluate their ability to highlight the presence of a specific type of bias in the data.

5.1. Methods

We propose three measures to detect the presence of each type of bias in the data. It is worth noting that in the previous section, we used a biased training set to train algorithms and an unbiased test set for their evaluation. In this section, we take the perspective of practitioners looking to evaluate their development dataset σ for biases without necessarily having access to unbiased data sources. We therefore split σ into identically distributed training and test sets.

Underrepresentation. We propose the Representation Difference (RD), to measure the underrepresentation of the disadvantaged group.

$$\mathrm{RD}(\sigma) = \frac{|\sigma_a| - |\sigma_d|}{|\sigma|} = \Pr_{\sigma}(s = a) - \Pr_{\sigma}(s = d) \tag{10}$$

RD quantifies the difference between the prevalence of the advantaged and disadvantaged groups. RD is a directional measure: positive (negative) values indicate a larger proportion of individuals from the (dis)advantaged group. A group s can be considered fairly represented in σ if $\mathrm{RD}(\sigma)$ is within certain limits. For example, practitioners can pick a threshold based on the prevalence of $s = d$ in a target population.

Label Bias. We introduce two measures of systematic label noise against the disadvantaged group. Specifically, we train a base classifier $\hat{y} = g(x)$ and evaluate AUC curves distinguishing between sensitive groups. We let $p_g(x)$ denote the posterior probabilities obtained by the classifier and we define the cross-dataset AUC of $g(x)$ as

$$\mathrm{xAUC}_g(\sigma_1, \sigma_2) = \Pr(p_g(x_i) > p_g(x_j) | y_i = \oplus, y_j = \ominus, i \in \sigma_1, j \in \sigma_2),$$

i.e. the probability that $g(x)$ correctly ranks a positive item from σ_1 higher than a negative one from σ_2.

Our first measure, initially introduced by Kallus and Zhou (2019), leverages a partition of σ into a disadvantaged set σ_d and an advantaged set σ_a, computes both measures of cross-dataset AUC, and defines their difference as

$$\Delta\mathrm{xAUC}_{\sigma} = \mathrm{xAUC}(\sigma_a, \sigma_d) - \mathrm{xAUC}(\sigma_d, \sigma_a). \tag{11}$$

[4] $h(x)$ takes as input the same features x we used to train $g(x)$.

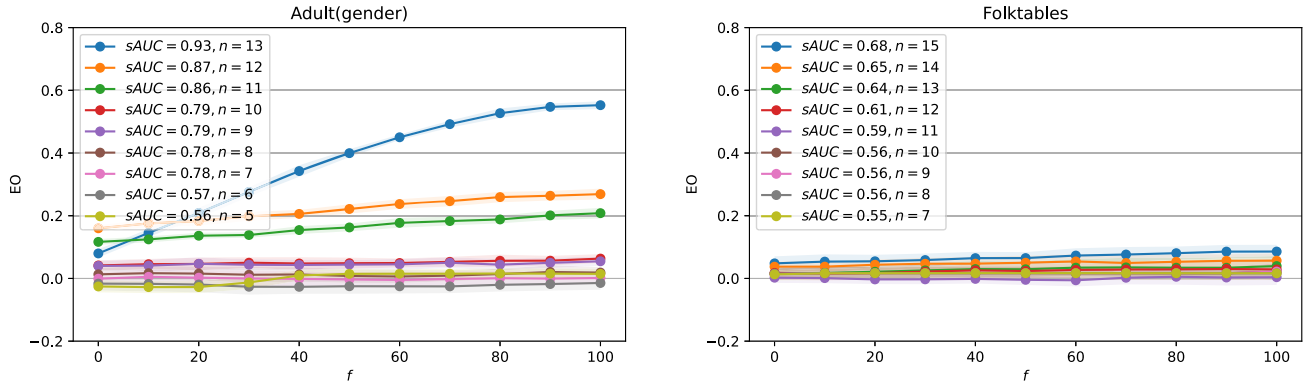

Fig. 3. Proxies exacerbate the risk of algorithmic discrimination caused by label bias. EO (y axis) increases with label bias f (x axis). This effect is mediated by proxies: weaker proxies (lower sAUC) correspond to a lower slope and a weaker effect of label bias on fairness.

Positive values indicate that pairs of advantaged positives and disadvantaged negatives are easier to separate correctly than pairs of disadvantaged positives and advantaged negatives.

Next, we define the within-group AUC difference as

$$\begin{aligned}\Delta\text{wAUC}_\sigma &= \text{AUC}_{\sigma_a}(g) - \text{AUC}_{\sigma_d}(g)\\ &= \Pr_{\sigma_a}(p_g(x_i) > p_g(x_j) \mid y_i = \oplus, y_j = \ominus)\\ &\quad - \Pr_{\sigma_d}(p_g(x_i) > p_g(x_j) \mid y_i = \oplus, y_j = \ominus),\end{aligned} \tag{12}$$

i.e. we compute the AUC for advantaged ($s = a$) and disadvantaged items ($s = d$) separately, and measure their difference. Large absolute values indicate a better separability for one group. Notice that both Eqs. (11) and (12) are directional: positive (negative) values indicate better separability for the (dis)advantaged group.

Finally, we compute Separation Difference (SD) as their average

$$\text{SD}(\sigma) = \frac{\Delta\text{xAUC}_\sigma + \Delta\text{wAUC}_\sigma}{2} \tag{13}$$

and employ it in the remainder of this section. We expect label bias to worsen the separability for the disadvantaged group and therefore yield high values of SD.

Proxies. To quantify the information about protected features encoded in non-protected ones, which may act as proxies, we train a classifier $h(\cdot) : x \to s$ to predict sensitive attributes from non-sensitive ones. The classifier's performance is then evaluated using the area under the ROC curve for the s predictor (sAUC).

$$s\text{AUC}(\sigma) = \text{AUC}_\sigma(h) \tag{14}$$

Higher values of this metric indicate a better ability of the classifier to predict the sensitive feature from the non-sensitive one, indicating stronger proxies.

5.2. Experiments

Setup. We leverage bias injection mechanisms to test bias detection. We use part of σ to train a classifier and part of σ to evaluate its performance. We maintain an 80-10-10 train-validation-test split for tabular datasets and NIH and a 70-15-15 split for Fitzpatrick17k. As in Section 4, we subsample the disadvantaged group by varying the underrepresentation factor $u \in \{0, 0.2, 0.8, 1\}$.

Similarly, we inject label bias letting the flip factor vary in $f \in \{0, 0.2, 0.8, 1\}$.[5]

Finally, we inject proxies through the additive mechanism presented in Section 3.4. We expand the input space with an additional feature derived as the sum of the sensitive attribute s and a normal variable with zero mean and decreasing standard deviation.[6] The standard deviation varies to achieve Pearson correlation coefficients between the sensitive attribute and the additional feature of approximately $\{0, 0.25, 0.50, 0.75, 1\}$. A correlation of 0 corresponds to the baseline scenario with no additional feature, while a correlation of 1 reflects the maximally biased scenario in which the additional feature is identical to the protected attribute.

Results. In Fig. 4, we present bias detection results for Folktables and NIH. Experimental results for the other datasets are provided in Appendix C. As anticipated, each bias detection metric specifically captures the type of bias it is designed to identify. The metrics exhibit strong variation in the diagonal panels of Fig. 4, while remaining relatively stable across the remaining panels.

Specifically, underrepresentation is suitably captured by RD increasing linearly in the first column, while SD and sAUC exhibit minor oscillations in their average values. Notice that extreme underrepresentation leads to large variations around mean values for SD and sAUC due to numerical instability (see Footnote 5). In the second column, label bias leads to an increase in SD, while RD and sAUC remain constant. Finally, proxies of increasing strength are detected by sAUC in the third column, while RD and SD remain stable. These trends are consistent across nearly all datasets, except for the SD metric on Compas, whose increase with f is barely noticeable. Overall, these results show the effectiveness of the proposed measures in detecting specific data biases.

Interpretation. After confirming the influence of data bias on algorithmic fairness in the previous section, in this section, we have provided a demonstration of bias detection based on principled measures. Each measure can detect a specific type of data bias. Crucially, their computation requires no access to unbiased test sets, making them widely applicable in practice.

5.3. Data bias profile

Based on these measures, we initiate the development of *Data Bias Profiles* (DBP), an extensible quantitative framework to describe data bias. We envision that the DBP will be used in fairness work to highlight biases that can lead to discrimination and inform decisions on fairness-enhancing interventions. Additionally, DBP is suited to summarize biases in the documentation accompanying a dataset. We position this as an initial but foundational contribution—significant work remains to validate, refine, and extend the DBP into a mature, fully realized

[5] Since extreme values such as $u = 1$ and $f = 1$ would render computation of SD and sAUC infeasible, we replace them with $u = 0.95$ and $f = 0.95$. It is worth reiterating that, in this section, training sets and test sets are drawn from the same distribution, differently from the previous section, where test sets were unbiased.

[6] For image datasets, the additional proxy is fed to the penultimate layer of the neural network.

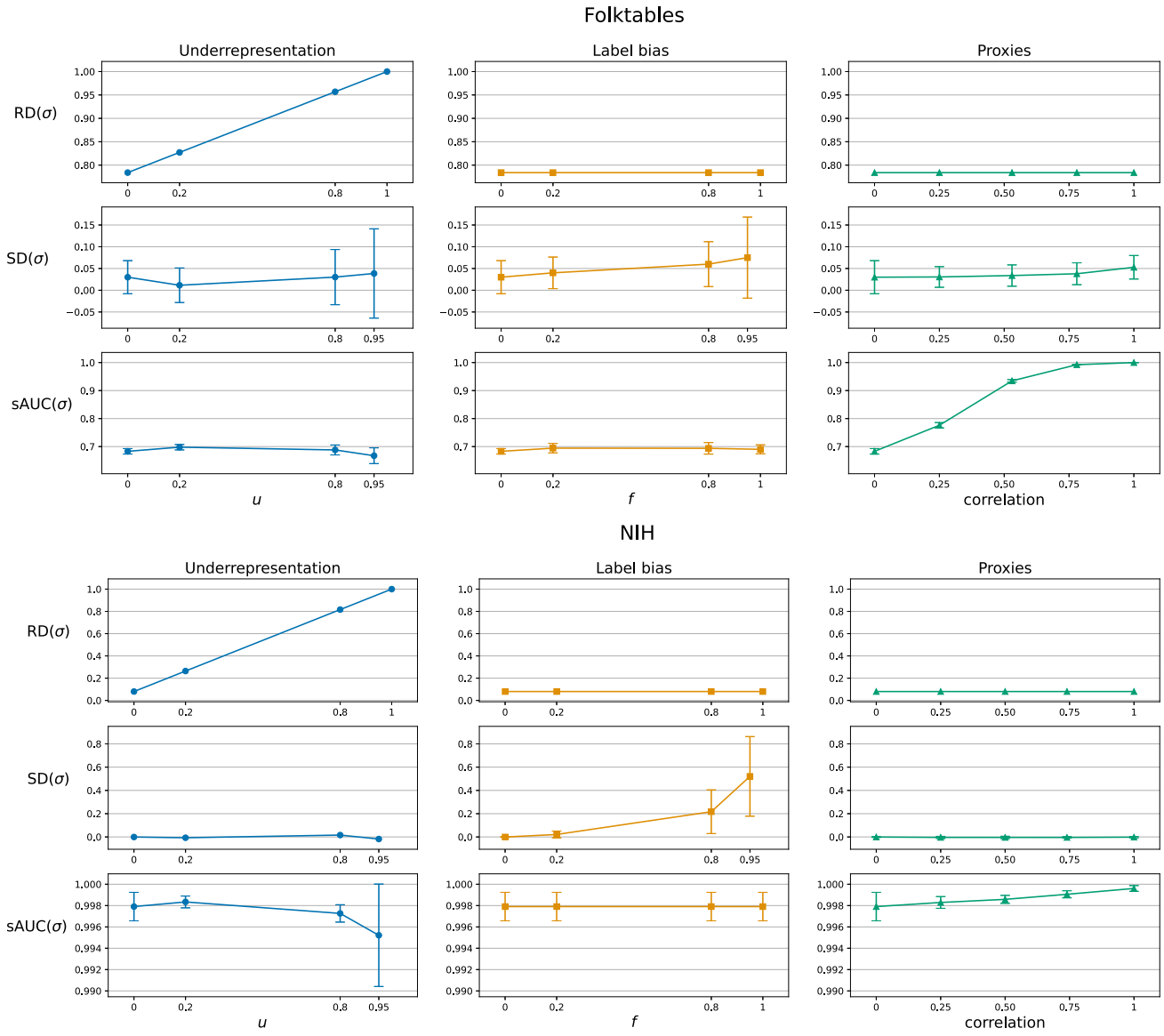

Fig. 4. The proposed measures capture specific types of bias. Bias detection on Folktables and NIH. Columns correspond to three bias injection mechanisms; rows correspond to bias detection measures. Measures vary when the corresponding bias increases (diagonal) and remain relatively flat with other biases (off-diagonal).

tool. Achieving this vision will require a coordinated effort across the research community to evolve the DBP into a robust and widely adopted quantitative framework.

Fig. 5 demonstrates DBP with a practical use case. On the left, we present DBPs for two datasets. Adult (marital-status) presents considerable label bias and strong proxies. As shown in Section 4, this entails a high probability of algorithmic discrimination. Folktables, on the other hand, has low label bias and weak proxies, highlighting a contained risk of algorithmic discrimination. This is confirmed by the right panel in Fig. 5, where round markers depict the average unfairness (DP and EO – Eqs. (7) and (8)) for a logistic regression model over ten random splits of the datasets. DBPs also hint at the effectiveness of proxy reduction as a bias mitigation strategy. Tackling the strong proxies in Adult can largely reduce unfairness; Folktables, on the other hand, displays weak proxies and is unlikely to benefit from the same approach. We test this hypothesis by removing from each dataset the feature that is most strongly correlated with the sensitive attribute. Star-shaped markers depict EO and DP for the resulting models in the right panel of Fig. 5. As predicted, proxy removal strongly curbs unfairness on Adult (marital-status), while its effect on Folktables is barely noticeable.

6. Discussion

We discuss our results in the broader context of responsible AI. Table 7 summarizes the implications of this work for researchers and practitioners.

Underrepresentation in training is overemphasized. Increasing the prevalence of vulnerable groups in training sets is touted as the key strategy to achieve fairness. In stark contradiction with this credence, Section 4 shows that extreme variations in the prevalence of protected groups have a minor impact on fairness across diverse datasets, machine learning models, and metric choices. *First and foremost, we urge practitioners and researchers against using these results as a blanket justification to neglect inclusion efforts.* Although challenging, expanding and diversifying datasets in a responsible manner is fundamental for keeping algorithmic systems in check. We provide a more nuanced interpretation. High-quality data from disadvantaged groups annotated with sensitive attributes is likely to be scarce, stemming e.g. from targeted curation efforts. Since including disadvantaged groups in training sets is often unlikely to bring meaningful improvements, we recommend prioritizing this data for reliable system evaluations (rather than training), including

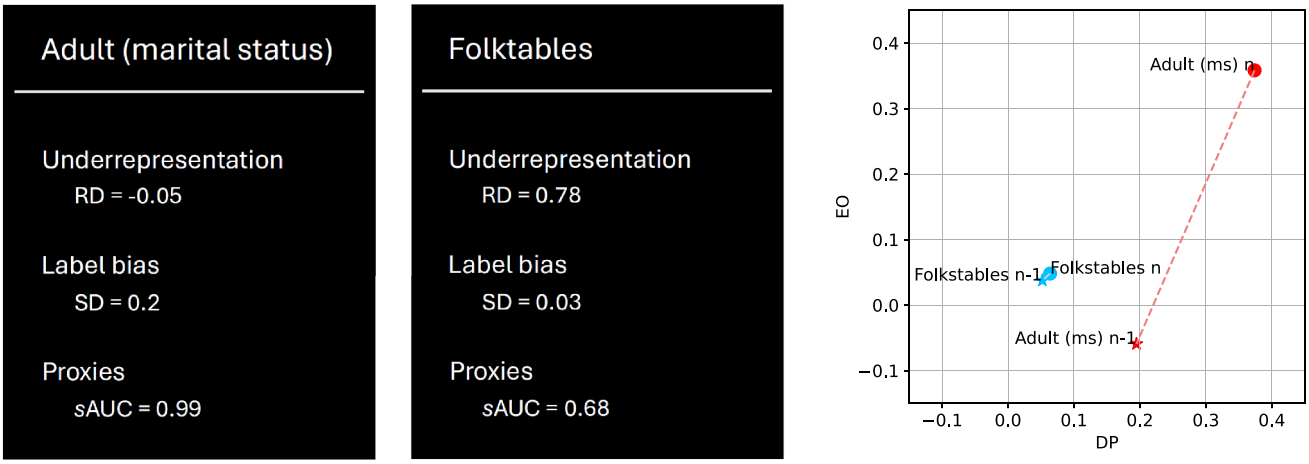

Fig. 5. Data bias profiles hint at the risk of algorithmic discrimination and effectiveness of fairness intervention. DBP of Adult (left) and Folktables (center); on the right, model fairness summarized by demographic parity (x axis) and equality of opportunity (y). DBP highlights strong proxies and label bias for Adult (marital status), leading to a high risk of discrimination (red round marker), which can be mitigated with proxy reduction (star-shaped marker). Folktables has weak proxies and label bias, translating into lower unfairness and ineffectiveness of proxy mitigation (blue markers).

Table 7
Implications. Takeaways and recommendations for algorithmic fairness researchers (R) and practitioners (P).

Takeaway	Recommendations
Underrepresentation in training is overemphasized	(R) Critically reconsider widespread belief (P) Use scarce annotated data for reliable evaluations
Label bias is critical	(R) Research techniques for label bias detection (P) Seek and critically assess multiple target labels
Data Bias Profiles (DBP) can link data bias & group fairness	(R) Diversify fairness testbeds with DBP (P) Use DBP to select fairness interventions (P) Include DBP in data documentation

measurements of algorithmic fairness. If fairness evaluations yield problematic results, practitioners should carry out an assessment of multiple bias factors that go beyond underrepresentation.

Label bias is critical. Systematic bias against vulnerable groups in target variables (in short: label bias) is a common occurrence due to structural societal differences. Section 4 shows that label bias has a major impact on fairness across diverse metrics, models, and datasets as well as significant interactions with other types of data bias. For example, in the presence of label bias, increasing training set representation can have a detrimental effect on vulnerable groups. In this setting, underrepresentation may actually benefit vulnerable groups, challenging conventional wisdom. Notice that label bias is especially critical also because there is a high risk it will go undetected if models are trained and evaluated with datasets that exhibit the same bias (e.g. identically distributed training and test sets). Based on these findings, we make two recommendations. Practitioners should seek multiple target variables for their models and carefully choose the most suitable one(s) to minimize the potential for label bias. This effort should blend qualitative approaches grounded in domain expertise with quantitative approaches for bias detection. In concert, researchers should develop reliable techniques for label bias detection. Section 5 is a first step in this direction.

Data Bias Profiles (DBP) can link fairness and bias. Overall, this work contributes a list of independent data biases with mechanisms to study them (Section 3), a study of their joint influence on fairness (Section 4), and principled methods for bias quantification (Section 5). Building upon these contributions, we advocate a broader community effort to develop the Data Bias Profile (DBP), as a first attempt to summarize key bias indicators in a unified format. Rather than presenting a finalized framework, we offer an extensible prototype. In their current form, DBPs are brief summaries of datasets that leverage bias quantification methods for a principled analysis of fairness problems guided by data. Model developers should use DBPs to reason about sources of model unfairness and select tailored approaches for mitigation. For example, detecting strong label bias may direct a developer towards fairness interventions for target label repair. Additionally, DBPs can develop into reference documentation frameworks that practitioners will use to comply with data governance requirements, including bias detection provisions specified in the AI Act. From a research standpoint, the DBP offers promising directions. DBPs can guide the development of fairness benchmarks, i.e. standardized collections to evaluate alternative fairness algorithms. For example, datasets can be distinguished based on the presence of strong proxies and weak proxies.

As we have shown, both vanilla algorithms (solely focused on accuracy) and fairness interventions behave differently based on the strength of proxies encoded in non-sensitive features. Fairness testbeds should thus include both strong-proxy and weak-proxy datasets to evaluate models under diverse conditions.

Finally, DBPs may bridge hundreds of fairness algorithms (Hort, Chen, Zhang, Harman, & Sarro, 2024) and datasets (Fabris et al., 2022), by helping to answer the key question: given a model that produces unfair predictions on a dataset, which type of fairness-enhancing algorithm is most suitable for that type of data and algorithm?

6.1. Limitations

This study has some limitations. First, we only cover three types of data bias. Although these are the most cited in scholarly articles and technical reports, different types of data bias are possible (Baumann et al., 2023; Mehrabi et al., 2022). Future work should consider additional biases, including feature bias, omitted variable bias, and concept drift across protected groups. Second, we consider binary protected attributes. While most results generalize to multi-group attributes by casting them as one-vs-all problems, this may become impractical for large cardinality $|S|$. Natively catering to multi-group attributes will require careful adaptation. Third, we provide no thresholds to distinguish between mild and excessive bias with our detection mechanisms. In its current form, the DBP is useful for relative comparisons across datasets;

it will need further refinement to support thresholding. Fourth, while we experiment with popular and diverse fairness datasets, this is unlikely to be exhaustive of all settings encountered in practice. Future work should include additional datasets, with special attention to datasets with complementary properties. Finally, we note that it may be impossible to distinguish proper data bias, i.e. a shift between the data and a target population, from situations where the data is "uncorrupted", yet naturally encodes groupwise differences. Our work contributes a principled way to link numerical data properties with algorithmic fairness properties. Arguments on the source of those numerical properties are extremely valuable and complement our contributions.

7. Conclusion

Data biases are key drivers of algorithmic discrimination. While this fact is broadly recognized, their relative importance and interaction remain understudied. Our work targets this gap with a systematic study of bias conducive factors, their influence on algorithmic discrimination, and their detection through dedicated mechanisms. These are necessary steps to develop a shared lexicon to describe data bias, document it unambiguously, and link it to fairness interventions in a principled fashion.

To realize these goals, we call for a community-wide effort to expand, formalize, and critically assess the Data Bias Profile, paving the way for a shared and trustworthy approach to quantitative bias documentation.

This line of work will be critical to steer anti-discrimination policy toward technically meaningful standards and to translate algorithmic fairness research into law-abiding practice.

CRediT authorship contribution statement

Marina Ceccon: Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Methodology, Software; **Giandomenico Cornacchia:** Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Methodology; **Davide Dalle Pezze:** Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Methodology; **Alessandro Fabris:** Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Methodology, Supervision; **Gian Antonio Susto:** Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Methodology, Supervision.

Data availability

Data will be made available on request.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Marina Ceccon and Gian Antonio Susto reports financial support was provided by Ministry of Education and Merit. Alessandro Fabris reports financial support was provided by Alexander von Humboldt Foundation and the FINDHR project. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Alexander von Humboldt Foundation, the FINDHR project (Horizon Europe grant agreement ID: 101070212) (A. Fabris) and by Ministero dell'Università della Ricerca (MUR), iniziativa Dottorati PON (M. Ceccon, G.A. Susto).

Appendix A. Datasets

In this appendix, we present algorithmic fairness datasets and their processing in this work. Sensitive features (s) that are used for fairness evaluations are excluded from input features.

Adult[7] is a prominent dataset hosted by the UCI Machine Learning Repository, originally derived from the 1994 US Census database (Kohavi, 1996). The primary objective of this dataset is to predict whether an individual's annual income exceeds $50,000. We follow Donini, Oneto, Ben-David, Shawe-Taylor, and Pontil (2018), keeping all the features in the dataset. We use gender and marital-status as sensitive attributes.

Compas[8] comprises data from the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, a commercial tool used to predict recidivism among convicted individuals. The dataset, collected by ProPublica to audit the COMPAS system, surfaced discrimination against African-American defendants. We follow the preprocessing from Ruoss, Balunovic, Fischer, and Vechev (2020) and use the following variables.

- race (s): race of the defendants.
- age: age of the defendants.
- c_charge_degree: charge degree (F: Felony, M: Misdemeanor).
- diff_custody: time spent in custody.
- diff_jail: time spent in jail.
- sex: sex of the defendants.
- priors_count: number of prior criminal records.
- length_of_stay: number of days spent in jail.
- v_score_text: COMPAS quantized score, summarizing additional features used by the model but unavailable in the data collected by ProPulica.

The target variable is two_year_recid, indicating whether an individual re-offended within 2 years of being released. For protected attributes, we focus on race, distinguishing between Caucasian (advantaged group) and African-American defendants (disadvantaged group).

Crime[9] is a real-world dataset from the UCI Machine Learning Repository, focused on predicting violent crime rates across various communities in the US. The task involves predicting whether a community can be classified as violent based on its crime rates, specifically when the number of crimes exceeds the median value of crimes across all states. We follow the setup from Balunovic, Ruoss, and Vechev (2022), including binarized race as a sensitive attribute. We keep all the non-sensitive features for inference ($n = 127$).

Folktables[10] is a Python package designed to provide access to datasets derived from the US Census Bureau's American Community Survey (ACS) (Ding, Hardt, Miller, & Schmidt, 2021). The complete data underlying the folktables dataset comprises the full ACS census data, spanning all US states, multiple years, and prediction targets. In this work, we focus on the employment prediction task (ACSEmployment), filtering the data to include individuals aged between 16 and 90. We subsample at 1 % of the dataset size, stratifying on the target and sensitive label to maintain the distribution of the original data. We use the standard data loader keeping the following features:

- ESR: employment status of the individual, represented as a binary categorical feature (1: Employed, 0: Otherwise).
- RAC1P (s): detailed race recode (categorical values 1–9).
- AGEP: age in years, with a maximum value of 99.
- ANC: ancestry recode (categorical).
- CIT: citizenship status of the individual, represented as a categorical string.

[7] https://archive.ics.uci.edu/dataset/2/adult
[8] https://github.com/propublica/compas-analysis
[9] https://archive.ics.uci.edu/dataset/183/communities+and+crime
[10] https://github.com/socialfoundations/folktables

- DEAR: hearing difficulty (binary).
- DEYE: vision difficulty (binary).
- DIS: disability recode (1: citizen with disability, 2: without).
- DREM: cognitive difficulty of the individual, indicating if they have difficulty remembering, concentrating, or making decisions (binary).
- ESP: employment status of parents (categorical)
- MAR: marital status of the individual (categorical).
- MIG: Mobility status, indicating residence one year ago (categorical).
- MIL: military service (categorical).
- NATIVITY: binary variable indicating US native or foreign-born.
- RELP: relationship (categorical values 1–17).
- SCHL: educational attainment (categorical values 1–24, or NA).
- SEX: sex/Gender (1: Male, 2: Female).

The Employment Status Recode (ESR) is the target variable (equal to 1 if employed, 0 otherwise). The advantaged group consists of individuals with RAC1P equal to 1 (Caucasian), while the disadvantaged group includes all individuals with RAC1P values other than 1 (other races).

German[11] is another widely recognized dataset from the UCI Machine Learning Repository, encompassing records of bank loan applications in Germany. This dataset contains demographic and financial details of applicants, along with the loan approval outcomes. The primary prediction task is binary, aimed at determining creditworthiness based on loan repayment. We use the following features:

- age (s): The age of the applicant, binarized with a 25-year threshold.
- amount: credit amount in Deutsche Mark.
- credit_history: history of credit usage and repayment by the applicant (categorical).
- duration: duration of the loan in months (numeric).
- employment_duration: tenure with current employer (numeric).
- housing: type of housing (categorical).
- installment_rate: percentage of applicant's income allocated to loan installments (categorical).
- job: applicant's job and employability (categorical).
- number_credits: number of credits with this bank (categorical).
- other_debtors: indication of an additional debtor or a guarantor for the credit (categorical).
- other_installment_plans: installment plans with other banks (categorical).
- people_liable: number of people who are financially dependent on the applicant (categorical).
- property: applicant's most valuable property (categorical).
- purpose: purpose of loan (categorical).
- present_residence: years lived at current address (categorical).
- savings: savings account balance (categorical).
- status: status of the individual's saving accounts (categorical).
- telephone: whether the applicant has a registered telephone line.

ChestX-ray14 (NIH)[12] is a comprehensive medical imaging dataset containing 112,120 frontal-view chest X-ray images from 30,805 unique patients, collected between 1992 and 2015 (Wang et al., 2017). Disease labels for fourteen common thoracic conditions were extracted from radiological reports using natural language processing. The labeled conditions include: atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, mas, nodule, pneumonia, pneumothorax. The associated classification task is multi-label, with each of the 14 target labels indicating the presence of a specific disease. Additionally, patient metadata provides information on both gender and age.

We include only one image per patient, as previous studies have shown that this approach reduces bias without significantly affecting overall performance (Weng, Bigdeli, Petersen, & Feragen, 2023). We conduct evaluations independently for each disease and report macro-averaged metrics across all diseases. In this study, binary gender is the sensitive attribute.

Fitzpatrick17k[13] is a medical imaging dataset containing 16,577 clinical images (Groh et al., 2021), each annotated with labels for skin conditions and skin type based on the Fitzpatrick scale (Fitzpatrick, 1988). The images were sourced from two open-access dermatology atlases: 12,672 images from DermaAmin and 3905 from Atlas Dermatologico.[14] The dataset includes 114 distinct disease labels and two additional levels of aggregated skin condition classifications, structured according to the skin lesion taxonomy proposed by Esteva et al. (2017). At the broadest classification level, skin conditions are divided into three main categories: benign lesions, malignant lesions and non-neoplastic lesions. In a more detailed classification, skin conditions are categorized into nine types: inflammatory, malignant epidermal, genodermatoses, benign dermal, benign epidermal, malignant melanoma, benign melanocyte, malignant cutaneous lymphoma, malignant dermal. Our classification task is based on a binary variable derived from the highest-level skin condition classification, distinguishing between neoplastic, hence tumoral (either benign or malignant), and non-neoplastic diseases. This mimics a preliminary assessment for the presence of tumoral conditions through assistive technology used by dermatology experts. The Fitzpatrick skin type labels follow a six-point scale, with 1 being the lightest and 6 the darkest skin type. We binarize them into light (1–4) and very dark (5-6).

Appendix B. Additional results on the effect of data bias

In this section, we report the effect of data bias on all machine learning models, datasets, and metrics. Specifically, performance is evaluated through balanced accuracy (Tables B.8 and B.14), while fairness is assessed through prediction quality parity (PQP – Tables B.9 and B.15), equal opportunity (EO – Tables B.12 and B.18), and demographic parity (DP – Tables B.13 and B.19). We zoom in on EO by breaking it down into groupwise true positive rate (TPR – Tables B.10, B.11, B.16, and B.17) components. We provide results for underrepresentation (Appendix B.1) and label bias (Appendix B.2).

As in Section 4, the tables include indicators of statistical significance; symbols (*) and (**) denote statistically significant differences with respect to the unbiased scenario thresholds at $p = .05$ and $p = .01$, respectively.

B.1. Underrepresentation

B.2. Label bias

[11] https://archive.ics.uci.edu/dataset/522/south+german+credit
[12] https://nihcc.app.box.com/v/ChestXray-NIHCC
[13] https://github.com/mattgroh/fitzpatrick17k
[14] https://atlasdermatologico.com.br

Table B.8
Balanced accuracy varying the percentage of minority-group points retained in the training set from $u = 0$ (no bias) to $u = 1$ (maximum bias).

			Balanced Accuracy			
Dataset	sensitive	model	$u = 0$ (no bias)	$u = 0.2$	$u = 0.8$	$u = 1$ (max bias)
Adult	gender	LR	0.77 ± 0.00	0.77 ± 0.00	0.77 ± 0.00	0.76 ± 0.01*
		RF	0.78 ± 0.00	0.78 ± 0.01	0.78 ± 0.00	0.77 ± 0.01*
		SVC	0.77 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.76 ± 0.01
	marital-status	LR	0.77 ± 0.00	0.77 ± 0.00	0.77 ± 0.00	0.77 ± 0.00
		RF	0.78 ± 0.00	0.78 ± 0.00	0.78 ± 0.00	0.78 ± 0.01
		SVC	0.77 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.77 ± 0.01
Compas	race	LR	0.67 ± 0.02	0.67 ± 0.02	0.67 ± 0.01	0.67 ± 0.01
		RF	0.69 ± 0.02	0.69 ± 0.01	0.69 ± 0.02	0.68 ± 0.02
		SVC	0.65 ± 0.03	0.65 ± 0.03	0.65 ± 0.02	0.65 ± 0.01
Crime	race	LR	0.84 ± 0.02	0.84 ± 0.02	0.83 ± 0.02	0.83 ± 0.02
		RF	0.83 ± 0.03	0.84 ± 0.03	0.83 ± 0.02	0.81 ± 0.03
		SVC	0.84 ± 0.02	0.84 ± 0.02	0.83 ± 0.02	0.83 ± 0.03
Folktables	race	LR	0.73 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	0.73 ± 0.01
		RF	0.78 ± 0.00	0.78 ± 0.01	0.77 ± 0.01*	0.78 ± 0.00
		SVC	0.72 ± 0.01	0.72 ± 0.01	0.72 ± 0.01	0.72 ± 0.01
German	age	LR	0.66 ± 0.06	0.65 ± 0.05	0.64 ± 0.07	0.65 ± 0.05
		RF	0.68 ± 0.06	0.68 ± 0.05	0.65 ± 0.05	0.64 ± 0.05
		SVC	0.66 ± 0.07	0.66 ± 0.07	0.63 ± 0.06	0.63 ± 0.06
NIH	gender	DenseNet	0.64 ± 0.01	0.64 ± 0.02	0.64 ± 0.02	0.63 ± 0.02
Fitzpatrick17k	skin type	vgg16	0.73 ± 0.01	0.70 ± 0.01**	0.73 ± 0.02	0.70 ± 0.02**

Table B.9
Prediction quality parity (PQP) varying the percentage of minority-group points retained in the training set from $u = 0$ (no bias) to $u = 1$ (maximum bias).

			PQP			
Dataset	sensitive	model	$u = 0$ (no bias)	$u = 0.2$	$u = 0.8$	$u = 1$ (max bias)
Adult	gender	LR	0.00 ± 0.01	0.00 ± 0.01	0.01 ± 0.01	0.06 ± 0.03**
		RF	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.01	0.11 ± 0.01**
		SVC	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.08 ± 0.01**
	marital-status	LR	0.09 ± 0.01	0.09 ± 0.02	0.09 ± 0.02	0.06 ± 0.02**
		RF	0.08 ± 0.01	0.08 ± 0.01	0.09 ± 0.01	0.02 ± 0.02**
		SVC	0.08 ± 0.02	0.09 ± 0.02	0.08 ± 0.02	0.06 ± 0.01*
Compas	race	LR	−0.06 ± 0.03	−0.06 ± 0.03	−0.06 ± 0.04	−0.06 ± 0.04
		RF	−0.01 ± 0.05	−0.02 ± 0.06	−0.02 ± 0.05	−0.02 ± 0.04
		SVC	−0.05 ± 0.03	−0.05 ± 0.04	−0.06 ± 0.03	−0.06 ± 0.04
Crime	race	LR	−0.02 ± 0.08	−0.01 ± 0.09	−0.04 ± 0.10	−0.06 ± 0.10
		RF	−0.03 ± 0.05	−0.04 ± 0.09	−0.04 ± 0.07	−0.08 ± 0.07
		SVC	−0.03 ± 0.08	0.01 ± 0.06	−0.05 ± 0.05	−0.06 ± 0.06
Folktables	race	LR	0.01 ± 0.04	0.01 ± 0.04	0.02 ± 0.04	0.02 ± 0.04
		RF	0.01 ± 0.03	0.00 ± 0.03	0.01 ± 0.03	0.01 ± 0.03
		SVC	0.02 ± 0.03	0.02 ± 0.03	0.02 ± 0.03	0.02 ± 0.03
German	age	LR	0.07 ± 0.13	0.05 ± 0.09	0.08 ± 0.08	0.08 ± 0.10
		RF	−0.01 ± 0.07	0.02 ± 0.07	0.06 ± 0.05	0.04 ± 0.06
		SVC	0.03 ± 0.10	0.03 ± 0.08	0.02 ± 0.10	0.02 ± 0.09
NIH	gender	DenseNet	−0.01 ± 0.01	0.00 ± 0.01	0.01 ± 0.01**	0.02 ± 0.01**
Fitzpatrick17k	skin type	vgg16	−0.01 ± 0.01	0.03 ± 0.01**	0.04 ± 0.02**	0.03 ± 0.03**

Table B.10
True positive rate (TPR) of the advantaged group varying the percentage of minority-group points retained in the training set from $u = 0$ (no bias) to $u = 1$ (maximum bias).

			TPR ($s = a$)			
Dataset	sensitive	model	$u = 0$ (no bias)	$u = 0.2$	$u = 0.8$	$u = 1$ (max bias)
Adult	gender	LR	0.61 ± 0.01	0.62 ± 0.01	0.62 ± 0.01	0.62 ± 0.01
		RF	0.64 ± 0.01	0.63 ± 0.01	0.64 ± 0.01	0.63 ± 0.01
		SVC	0.61 ± 0.01	0.61 ± 0.01	0.61 ± 0.01	0.61 ± 0.01
	marital-status	LR	0.65 ± 0.01	0.65 ± 0.01	0.65 ± 0.01	0.66 ± 0.01
		RF	0.66 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.67 ± 0.01
		SVC	0.64 ± 0.01	0.64 ± 0.01	0.65 ± 0.01	0.66 ± 0.01**
Compas	race	LR	0.85 ± 0.02	0.86 ± 0.01	0.87 ± 0.02	0.88 ± 0.02*
		RF	0.81 ± 0.03	0.81 ± 0.03	0.81 ± 0.02	0.81 ± 0.03
		SVC	0.85 ± 0.04	0.86 ± 0.05	0.91 ± 0.02**	0.92 ± 0.02**
Crime	race	LR	0.90 ± 0.04	0.90 ± 0.04	0.89 ± 0.04	0.91 ± 0.04
		RF	0.90 ± 0.03	0.90 ± 0.03	0.91 ± 0.03	0.92 ± 0.03
		SVC	0.91 ± 0.04	0.91 ± 0.03	0.91 ± 0.03	0.93 ± 0.03
Folktables	race	LR	0.83 ± 0.01	0.83 ± 0.01	0.83 ± 0.01	0.83 ± 0.01
		RF	0.85 ± 0.01	0.86 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
		SVC	0.85 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.85 ± 0.01
German	age	LR	0.89 ± 0.05	0.89 ± 0.05	0.91 ± 0.05	0.91 ± 0.05
		RF	0.93 ± 0.02	0.91 ± 0.04	0.93 ± 0.02	0.92 ± 0.06
		SVC	0.93 ± 0.02	0.92 ± 0.02	0.93 ± 0.04	0.94 ± 0.04
NIH	gender	DenseNet	0.46 ± 0.03	0.49 ± 0.02	0.45 ± 0.05	0.45 ± 0.04
Fitzpatrick17k	skin type	vgg16	0.70 ± 0.02	0.68 ± 0.02	0.68 ± 0.01*	0.69 ± 0.01

Table B.11
True positive rate (TPR) of the disadvantaged group varying the percentage of minority-group points retained in the training set from $u = 0$ (no bias) to $u = 1$ (maximum bias).

			TPR ($s = d$)			
Dataset	sensitive	model	$u = 0$ (no bias)	$u = 0.2$	$u = 0.8$	$u = 1$ (max bias)
Adult	gender	LR	0.53 ± 0.02	0.53 ± 0.02	0.53 ± 0.02	0.41 ± 0.07**
		RF	0.54 ± 0.02	0.54 ± 0.02	0.52 ± 0.02	0.34 ± 0.03**
		SVC	0.53 ± 0.02	0.53 ± 0.02	0.51 ± 0.03	0.36 ± 0.04**
	marital-status	LR	0.29 ± 0.03	0.29 ± 0.03	0.29 ± 0.03	0.37 ± 0.03**
		RF	0.35 ± 0.03	0.34 ± 0.03	0.33 ± 0.02	0.51 ± 0.04**
		SVC	0.30 ± 0.03	0.30 ± 0.03	0.31 ± 0.03	0.39 ± 0.02**
Compas	race	LR	0.68 ± 0.03	0.68 ± 0.03	0.71 ± 0.03	0.72 ± 0.02**
		RF	0.66 ± 0.03	0.67 ± 0.03	0.68 ± 0.04	0.70 ± 0.03**
		SVC	0.68 ± 0.06	0.69 ± 0.07	0.76 ± 0.05*	0.78 ± 0.03**
Crime	race	LR	0.57 ± 0.10	0.54 ± 0.12	0.59 ± 0.14	0.63 ± 0.13
		RF	0.57 ± 0.07	0.58 ± 0.10	0.62 ± 0.08	0.75 ± 0.09**
		SVC	0.59 ± 0.08	0.55 ± 0.08	0.61 ± 0.07	0.63 ± 0.09
Folktables	race	LR	0.79 ± 0.04	0.79 ± 0.04	0.78 ± 0.04	0.78 ± 0.05
		RF	0.83 ± 0.03	0.84 ± 0.04	0.85 ± 0.03	0.84 ± 0.04
		SVC	0.80 ± 0.04	0.81 ± 0.04	0.81 ± 0.04	0.81 ± 0.03
German	age	LR	0.82 ± 0.13	0.82 ± 0.12	0.84 ± 0.10	0.85 ± 0.09
		RF	0.90 ± 0.07	0.87 ± 0.07	0.90 ± 0.07	0.89 ± 0.07
		SVC	0.87 ± 0.09	0.89 ± 0.09	0.92 ± 0.10	0.90 ± 0.11
NIH	gender	DenseNet	0.44 ± 0.05	0.45 ± 0.03	0.42 ± 0.04	0.39 ± 0.03*
Fitzpatrick17k	skin type	vgg16	0.61 ± 0.05	0.58 ± 0.06	0.57 ± 0.05	0.58 ± 0.05

Table B.12
Equal opportunity (EO) varying the percentage of minority-group points retained in the training set from $u = 0$ (no bias) to $u = 1$ (maximum bias).

			EO			
Dataset	sensitive	model	$u = 0$ (no bias)	$u = 0.2$	$u = 0.8$	$u = 1$ (max bias)
Adult	gender	LR	0.08 ± 0.02	0.09 ± 0.02	0.09 ± 0.02	0.21 ± 0.07**
		RF	0.10 ± 0.02	0.09 ± 0.02	0.12 ± 0.03	0.30 ± 0.03**
		SVC	0.08 ± 0.02	0.08 ± 0.02	0.10 ± 0.03	0.25 ± 0.03**
	marital-status	LR	0.36 ± 0.03	0.36 ± 0.03	0.37 ± 0.03	0.29 ± 0.04**
		RF	0.31 ± 0.03	0.31 ± 0.03	0.33 ± 0.03	0.16 ± 0.05**
		SVC	0.34 ± 0.03	0.34 ± 0.03	0.34 ± 0.03	0.27 ± 0.02**
Compas	race	LR	0.17 ± 0.03	0.17 ± 0.03	0.16 ± 0.02	0.16 ± 0.02
		RF	0.15 ± 0.05	0.14 ± 0.05	0.13 ± 0.04	0.11 ± 0.05
		SVC	0.18 ± 0.04	0.17 ± 0.04	0.15 ± 0.04	0.13 ± 0.03**
Crime	race	LR	0.33 ± 0.11	0.36 ± 0.12	0.30 ± 0.14	0.28 ± 0.13
		RF	0.33 ± 0.06	0.32 ± 0.11	0.29 ± 0.08	0.16 ± 0.09**
		SVC	0.32 ± 0.08	0.36 ± 0.09	0.30 ± 0.07	0.31 ± 0.07
Folktables	race	LR	0.05 ± 0.04	0.05 ± 0.04	0.05 ± 0.04	0.05 ± 0.04
		RF	0.02 ± 0.04	0.02 ± 0.03	0.02 ± 0.03	0.02 ± 0.04
		SVC	0.04 ± 0.03	0.04 ± 0.04	0.04 ± 0.03	0.04 ± 0.03
German	age	LR	0.07 ± 0.13	0.06 ± 0.11	0.07 ± 0.09	0.06 ± 0.07
		RF	0.03 ± 0.06	0.03 ± 0.06	0.03 ± 0.07	0.03 ± 0.08
		SVC	0.06 ± 0.08	0.03 ± 0.08	0.01 ± 0.08	0.04 ± 0.08
NIH	gender	DenseNet	0.01 ± 0.02	0.04 ± 0.01**	0.03 ± 0.02	0.06 ± 0.02**
Fitzpatrick17k	skin type	vgg16	0.09 ± 0.05	0.11 ± 0.06	0.11 ± 0.05	0.11 ± 0.05

Table B.13
Demographic parity (DP) varying the percentage of minority-group points retained in the training set from $u = 0$ (no bias) to $u = 1$ (maximum bias).

			DP			
Dataset	sensitive	model	$u = 0$ (no bias)	$u = 0.2$	$u = 0.8$	$u = 1$ (max bias)
Adult	gender	LR	0.18 ± 0.01	0.18 ± 0.01	0.18 ± 0.01	0.20 ± 0.01**
		RF	0.18 ± 0.01	0.18 ± 0.01	0.18 ± 0.01	0.21 ± 0.01**
		SVC	0.17 ± 0.00	0.17 ± 0.00	0.18 ± 0.01*	0.20 ± 0.01**
	marital-status	LR	0.37 ± 0.01	0.37 ± 0.01	0.38 ± 0.01	0.36 ± 0.02
		RF	0.36 ± 0.02	0.35 ± 0.02	0.36 ± 0.01	0.31 ± 0.02**
		SVC	0.36 ± 0.01	0.36 ± 0.01	0.36 ± 0.01	0.34 ± 0.01**
Compas	race	LR	0.26 ± 0.03	0.27 ± 0.03	0.26 ± 0.03	0.25 ± 0.03
		RF	0.21 ± 0.05	0.20 ± 0.04	0.19 ± 0.04	0.17 ± 0.05
		SVC	0.26 ± 0.02	0.25 ± 0.02	0.23 ± 0.03**	0.22 ± 0.03**
Crime	race	LR	0.62 ± 0.06	0.63 ± 0.05	0.61 ± 0.05	0.61 ± 0.05
		RF	0.62 ± 0.07	0.63 ± 0.07	0.61 ± 0.06	0.53 ± 0.07**
		SVC	0.62 ± 0.07	0.62 ± 0.06	0.61 ± 0.05	0.63 ± 0.06
Folktables	race	LR	0.06 ± 0.03	0.06 ± 0.03	0.06 ± 0.03	0.06 ± 0.03
		RF	0.05 ± 0.03	0.05 ± 0.03	0.04 ± 0.04	0.04 ± 0.03
		SVC	0.06 ± 0.03	0.06 ± 0.03	0.05 ± 0.02	0.05 ± 0.02
German	age	LR	0.06 ± 0.13	0.07 ± 0.09	0.04 ± 0.09	0.04 ± 0.08
		RF	0.09 ± 0.08	0.07 ± 0.08	0.03 ± 0.09	0.04 ± 0.08
		SVC	0.09 ± 0.10	0.06 ± 0.09	0.04 ± 0.07	0.07 ± 0.04
NIH	gender	DenseNet	0.00 ± 0.00	0.01 ± 0.00**	0.01 ± 0.01*	0.01 ± 0.01*
Fitzpatrick17k	skin type	vgg16	0.15 ± 0.02	0.14 ± 0.02	0.11 ± 0.02**	0.09 ± 0.02**

Table B.14
Balanced accuracy as the percentage of flipped positives in the disadvantaged group varies from $f = 0$ (no bias) to $f = 1$ (maximum bias).

			Balanced Accuracy			
Dataset	sensitive	model	$f = 0$ (no bias)	$f = 0.2$	$f = 0.8$	$f = 1$ (max bias)
Adult	gender	LR	0.77 ± 0.00	0.76 ± 0.01*	0.73 ± 0.00**	0.73 ± 0.00**
		RF	0.78 ± 0.00	0.77 ± 0.01*	0.74 ± 0.01**	0.74 ± 0.00**
		SVC	0.77 ± 0.01	0.76 ± 0.01	0.73 ± 0.01**	0.73 ± 0.01**
	marital-status	LR	0.77 ± 0.00	0.76 ± 0.01*	0.75 ± 0.01**	0.75 ± 0.01**
		RF	0.78 ± 0.00	0.77 ± 0.00**	0.75 ± 0.00**	0.75 ± 0.00**
		SVC	0.77 ± 0.01	0.76 ± 0.01	0.75 ± 0.01**	0.75 ± 0.01**
Compas	race	LR	0.67 ± 0.02	0.68 ± 0.02	0.58 ± 0.01**	0.55 ± 0.01**
		RF	0.69 ± 0.02	0.69 ± 0.02	0.59 ± 0.01**	0.57 ± 0.01**
		SVC	0.65 ± 0.03	0.66 ± 0.02	0.51 ± 0.01**	0.50 ± 0.00**
Crime	race	LR	0.84 ± 0.02	0.83 ± 0.02	0.82 ± 0.02	0.81 ± 0.03
		RF	0.83 ± 0.03	0.83 ± 0.02	0.82 ± 0.03	0.81 ± 0.02
		SVC	0.84 ± 0.02	0.84 ± 0.02	0.82 ± 0.02	0.81 ± 0.02*
Folktables	race	LR	0.73 ± 0.01	0.73 ± 0.01	0.72 ± 0.01	0.72 ± 0.01
		RF	0.78 ± 0.00	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01
		SVC	0.72 ± 0.01	0.72 ± 0.01	0.73 ± 0.01	0.73 ± 0.01
German	age	LR	0.66 ± 0.06	0.68 ± 0.07	0.70 ± 0.08	0.70 ± 0.08
		RF	0.68 ± 0.06	0.67 ± 0.05	0.72 ± 0.05	0.71 ± 0.05
		SVC	0.66 ± 0.07	0.66 ± 0.07	0.70 ± 0.07	0.70 ± 0.06
NIH	gender	DenseNet	0.64 ± 0.02	0.65 ± 0.01	0.61 ± 0.02*	0.59 ± 0.02**
Fitzpatrick17k	skin type	vgg16	0.73 ± 0.01	0.71 ± 0.01**	0.70 ± 0.02**	0.70 ± 0.02**

Table B.15
Prediction quality parity (PQP) as the percentage of flipped positives in the disadvantaged group varies from $f = 0$ (no bias) to $f = 1$ (maximum bias).

			PQP			
Dataset	sensitive	model	$f = 0$ (no bias)	$f = 0.2$	$f = 0.8$	$f = 1$ (max bias)
Adult	gender	LR	0.00 ± 0.01	0.06 ± 0.01**	0.21 ± 0.01**	0.23 ± 0.01**
		RF	0.01 ± 0.01	0.06 ± 0.02**	0.24 ± 0.01**	0.25 ± 0.01**
		SVC	0.01 ± 0.01	0.05 ± 0.01**	0.22 ± 0.01**	0.24 ± 0.01**
	marital-status	LR	0.09 ± 0.01	0.12 ± 0.02**	0.21 ± 0.01**	0.21 ± 0.01**
		RF	0.08 ± 0.01	0.11 ± 0.02**	0.25 ± 0.01**	0.25 ± 0.01**
		SVC	0.08 ± 0.02	0.11 ± 0.01**	0.23 ± 0.01**	0.24 ± 0.01**
Compas	race	LR	−0.06 ± 0.03	−0.03 ± 0.05	0.06 ± 0.02**	0.05 ± 0.02**
		RF	−0.01 ± 0.05	0.00 ± 0.06	0.06 ± 0.03**	0.05 ± 0.02**
		SVC	−0.05 ± 0.03	−0.04 ± 0.05	0.01 ± 0.01**	0.00 ± 0.00**
Crime	race	LR	−0.02 ± 0.08	0.01 ± 0.07	0.11 ± 0.10*	0.13 ± 0.08**
		RF	−0.03 ± 0.05	0.02 ± 0.06	0.16 ± 0.07**	0.17 ± 0.06**
		SVC	−0.03 ± 0.08	0.01 ± 0.08	0.10 ± 0.07**	0.14 ± 0.07**
Folktables	race	LR	0.01 ± 0.04	0.01 ± 0.04	0.02 ± 0.04	0.02 ± 0.04
		RF	0.01 ± 0.03	0.00 ± 0.02	0.00 ± 0.03	0.01 ± 0.03
		SVC	0.02 ± 0.03	0.02 ± 0.04	0.02 ± 0.04	0.02 ± 0.04
German	age	LR	0.07 ± 0.13	0.06 ± 0.15	0.05 ± 0.16	0.06 ± 0.14
		RF	−0.01 ± 0.07	0.05 ± 0.09	0.07 ± 0.09	0.09 ± 0.14
		SVC	0.03 ± 0.10	0.02 ± 0.10	0.02 ± 0.11	0.03 ± 0.15
NIH	gender	DenseNet	−0.01 ± 0.01	0.01 ± 0.01**	0.10 ± 0.01**	0.14 ± 0.01**
Fitzpatrick17k	skin type	vgg16	−0.01 ± 0.01	0.01 ± 0.01**	0.20 ± 0.02**	0.05 ± 0.01**

Table B.16
True positive rate (TPR) of the advantaged group as the percentage of flipped positives in the disadvantaged group varies from $f = 0$ (no bias) to $f = 1$ (maximum bias).

			TPR ($s = a$)			
Dataset	sensitive	model	$f = 0$ (no bias)	$f = 0.2$	$f = 0.8$	$f = 1$ (max bias)
Adult	gender	LR	0.61 ± 0.01	0.61 ± 0.01	0.60 ± 0.01	0.60 ± 0.01
		RF	0.64 ± 0.01	0.63 ± 0.01	0.62 ± 0.01**	0.62 ± 0.01**
		SVC	0.61 ± 0.01	0.60 ± 0.01	0.60 ± 0.01	0.59 ± 0.01**
	marital-status	LR	0.65 ± 0.01	0.65 ± 0.01	0.65 ± 0.01	0.65 ± 0.01
		RF	0.66 ± 0.01	0.66 ± 0.01	0.65 ± 0.01	0.66 ± 0.01
		SVC	0.64 ± 0.01	0.64 ± 0.01	0.65 ± 0.01	0.65 ± 0.01
Compas	race	LR	0.85 ± 0.02	0.76 ± 0.03**	0.33 ± 0.04**	0.22 ± 0.04**
		RF	0.81 ± 0.03	0.72 ± 0.03**	0.34 ± 0.05**	0.25 ± 0.05**
		SVC	0.85 ± 0.04	0.79 ± 0.03**	0.03 ± 0.06**	0.00 ± 0.00**
Crime	race	LR	0.90 ± 0.04	0.89 ± 0.04	0.87 ± 0.04	0.85 ± 0.05
		RF	0.90 ± 0.03	0.89 ± 0.03	0.86 ± 0.03**	0.85 ± 0.02**
		SVC	0.91 ± 0.04	0.90 ± 0.04	0.88 ± 0.04	0.87 ± 0.04
Folktables	race	LR	0.83 ± 0.01	0.82 ± 0.01	0.77 ± 0.01**	0.75 ± 0.01**
		RF	0.85 ± 0.01	0.85 ± 0.01	0.83 ± 0.02**	0.83 ± 0.02**
		SVC	0.85 ± 0.01	0.83 ± 0.01**	0.78 ± 0.01**	0.76 ± 0.01**
German	age	LR	0.89 ± 0.05	0.88 ± 0.05	0.82 ± 0.05**	0.79 ± 0.06**
		RF	0.93 ± 0.02	0.91 ± 0.04	0.86 ± 0.05*	0.82 ± 0.05*
		SVC	0.93 ± 0.02	0.89 ± 0.06	0.83 ± 0.05**	0.82 ± 0.06**
NIH	gender	DenseNet	0.46 ± 0.03	0.49 ± 0.03	0.59 ± 0.06**	0.52 ± 0.05*
Fitzpatrick17k	skin type	vgg16	0.70 ± 0.02	0.70 ± 0.02	0.69 ± 0.02	0.70 ± 0.01

Table B.17
True positive rate (TPR) of the disadvantaged group as the percentage of flipped positives in the disadvantaged group varies from $f = 0$ (no bias) to $f = 1$ (maximum bias).

			TPR ($s = d$)			
Dataset	sensitive	model	$f = 0$ (no bias)	$f = 0.2$	$f = 0.8$	$f = 1$ (max bias)
Adult	gender	LR	0.53 ± 0.02	0.40 ± 0.03**	0.07 ± 0.02**	0.05 ± 0.02**
		RF	0.54 ± 0.02	0.43 ± 0.04**	0.06 ± 0.01**	0.04 ± 0.01**
		SVC	0.53 ± 0.02	0.42 ± 0.03**	0.06 ± 0.02**	0.03 ± 0.01**
	marital-status	LR	0.29 ± 0.03	0.22 ± 0.04**	0.03 ± 0.01**	0.02 ± 0.01**
		RF	0.35 ± 0.03	0.29 ± 0.03**	0.00 ± 0.00**	0.00 ± 0.00**
		SVC	0.30 ± 0.03	0.26 ± 0.03*	0.01 ± 0.01**	0.00 ± 0.00**
Compas	race	LR	0.68 ± 0.03	0.55 ± 0.05**	0.11 ± 0.02**	0.07 ± 0.02**
		RF	0.66 ± 0.03	0.57 ± 0.04**	0.16 ± 0.02**	0.09 ± 0.01**
		SVC	0.68 ± 0.06	0.58 ± 0.05**	0.01 ± 0.02**	0.00 ± 0.00**
Crime	race	LR	0.57 ± 0.10	0.50 ± 0.11	0.25 ± 0.15**	0.18 ± 0.12**
		RF	0.57 ± 0.07	0.47 ± 0.10	0.19 ± 0.12**	0.18 ± 0.10**
		SVC	0.59 ± 0.08	0.50 ± 0.11	0.26 ± 0.11**	0.18 ± 0.10**
Folktables	race	LR	0.79 ± 0.04	0.77 ± 0.04	0.69 ± 0.04**	0.66 ± 0.04**
		RF	0.83 ± 0.03	0.82 ± 0.03	0.77 ± 0.05*	0.77 ± 0.05*
		SVC	0.80 ± 0.04	0.78 ± 0.04	0.70 ± 0.04**	0.67 ± 0.04**
German	age	LR	0.82 ± 0.13	0.78 ± 0.15	0.60 ± 0.17*	0.54 ± 0.13**
		RF	0.90 ± 0.07	0.80 ± 0.09**	0.60 ± 0.10**	0.50 ± 0.16**
		SVC	0.87 ± 0.09	0.83 ± 0.14	0.61 ± 0.11**	0.55 ± 0.11**
NIH	gender	DenseNet	0.44 ± 0.05	0.40 ± 0.04	0.18 ± 0.07**	0.03 ± 0.01**
Fitzpatrick17k	skin type	vgg16	0.61 ± 0.05	0.53 ± 0.06*	0.48 ± 0.08**	0.42 ± 0.05**

Table B.18
Equal opportunity (EO) as the percentage of flipped positives in the disadvantaged group varies from $f = 0$ (no bias) to $f = 1$ (maximum bias).

			EO			
Dataset	sensitive	model	$f = 0$ (no bias)	$f = 0.2$	$f = 0.8$	$f = 1$ (max bias)
Adult	gender	LR	0.08 ± 0.02	0.21 ± 0.02**	0.52 ± 0.03**	0.55 ± 0.02**
		RF	0.10 ± 0.02	0.20 ± 0.04**	0.56 ± 0.01**	0.57 ± 0.02**
		SVC	0.08 ± 0.02	0.18 ± 0.02**	0.54 ± 0.02**	0.56 ± 0.01**
	marital-status	LR	0.36 ± 0.03	0.43 ± 0.04**	0.62 ± 0.02**	0.63 ± 0.02**
		RF	0.31 ± 0.03	0.37 ± 0.03**	0.65 ± 0.01**	0.66 ± 0.01**
		SVC	0.34 ± 0.03	0.39 ± 0.02**	0.65 ± 0.01**	0.65 ± 0.01**
Compas	race	LR	0.17 ± 0.03	0.21 ± 0.05	0.21 ± 0.05	0.15 ± 0.05
		RF	0.15 ± 0.05	0.15 ± 0.06	0.18 ± 0.05	0.15 ± 0.05
		SVC	0.18 ± 0.04	0.21 ± 0.04	0.02 ± 0.04**	0.00 ± 0.00**
Crime	race	LR	0.33 ± 0.11	0.39 ± 0.11	0.62 ± 0.17**	0.67 ± 0.15**
		RF	0.33 ± 0.06	0.42 ± 0.12	0.66 ± 0.13**	0.67 ± 0.10**
		SVC	0.32 ± 0.08	0.40 ± 0.12	0.62 ± 0.12**	0.69 ± 0.12**
Folktables	race	LR	0.05 ± 0.04	0.05 ± 0.04	0.08 ± 0.04	0.09 ± 0.04
		RF	0.02 ± 0.04	0.02 ± 0.03	0.05 ± 0.04	0.06 ± 0.03
		SVC	0.04 ± 0.03	0.06 ± 0.04	0.08 ± 0.03*	0.09 ± 0.04*
German	age	LR	0.07 ± 0.13	0.10 ± 0.14	0.22 ± 0.16	0.25 ± 0.10**
		RF	0.03 ± 0.06	0.11 ± 0.08	0.26 ± 0.09**	0.32 ± 0.15**
		SVC	0.06 ± 0.08	0.06 ± 0.11	0.22 ± 0.10**	0.26 ± 0.13**
NIH	gender	DenseNet	0.01 ± 0.02	0.09 ± 0.02**	0.40 ± 0.03**	0.48 ± 0.01**
Fitzpatrick17k	skin type	vgg16	0.09 ± 0.05	0.16 ± 0.06*	0.21 ± 0.08**	0.28 ± 0.02**

Table B.19
Demographic parity (DP) as the percentage of flipped positives in the disadvantaged group varies from $f = 0$ (no bias) to $f = 1$ (maximum bias).

			DP			
Dataset	sensitive	model	$f = 0$ (no bias)	$f = 0.2$	$f = 0.8$	$f = 1$ (max bias)
Adult	gender	LR	0.18 ± 0.01	0.20 ± 0.01**	0.25 ± 0.01**	0.25 ± 0.01**
		RF	0.18 ± 0.01	0.20 ± 0.01**	0.24 ± 0.01**	0.24 ± 0.01**
		SVC	0.17 ± 0.00	0.19 ± 0.01**	0.24 ± 0.01**	0.24 ± 0.01**
	marital-status	LR	0.37 ± 0.01	0.38 ± 0.01	0.40 ± 0.01**	0.40 ± 0.01**
		RF	0.36 ± 0.02	0.36 ± 0.02	0.38 ± 0.01*	0.38 ± 0.02
		SVC	0.36 ± 0.01	0.37 ± 0.01	0.39 ± 0.01**	0.39 ± 0.01**
Compas	race	LR	0.26 ± 0.03	0.28 ± 0.02	0.18 ± 0.04**	0.12 ± 0.03**
		RF	0.21 ± 0.05	0.20 ± 0.05	0.16 ± 0.04	0.12 ± 0.04**
		SVC	0.26 ± 0.02	0.28 ± 0.02	0.02 ± 0.03**	0.00 ± 0.00**
Crime	race	LR	0.62 ± 0.06	0.63 ± 0.06	0.69 ± 0.07	0.70 ± 0.08
		RF	0.62 ± 0.07	0.64 ± 0.07	0.67 ± 0.06	0.67 ± 0.06
		SVC	0.62 ± 0.07	0.64 ± 0.07	0.70 ± 0.06*	0.70 ± 0.07
Folktables	race	LR	0.06 ± 0.03	0.07 ± 0.03	0.09 ± 0.03	0.09 ± 0.03
		RF	0.05 ± 0.03	0.06 ± 0.03	0.09 ± 0.03*	0.09 ± 0.04
		SVC	0.06 ± 0.03	0.06 ± 0.03	0.09 ± 0.03	0.09 ± 0.03
German	age	LR	0.06 ± 0.13	0.11 ± 0.13	0.24 ± 0.16*	0.27 ± 0.15**
		RF	0.09 ± 0.08	0.12 ± 0.08	0.28 ± 0.12**	0.31 ± 0.11**
		SVC	0.09 ± 0.10	0.09 ± 0.11	0.26 ± 0.14*	0.30 ± 0.11**
NIH	gender	DenseNet	0.00 ± 0.00	0.02 ± 0.00**	0.10 ± 0.02**	0.10 ± 0.01**
Fitzpatrick17k	skin type	vgg16	0.15 ± 0.02	0.17 ± 0.02	0.22 ± 0.02**	0.23 ± 0.02**

Appendix C. Additional results on bias detection

In this section, we report bias detection results regarding all datasets except Folktables and NIH, which are discussed in Section 5 Figs. C.6–C.11

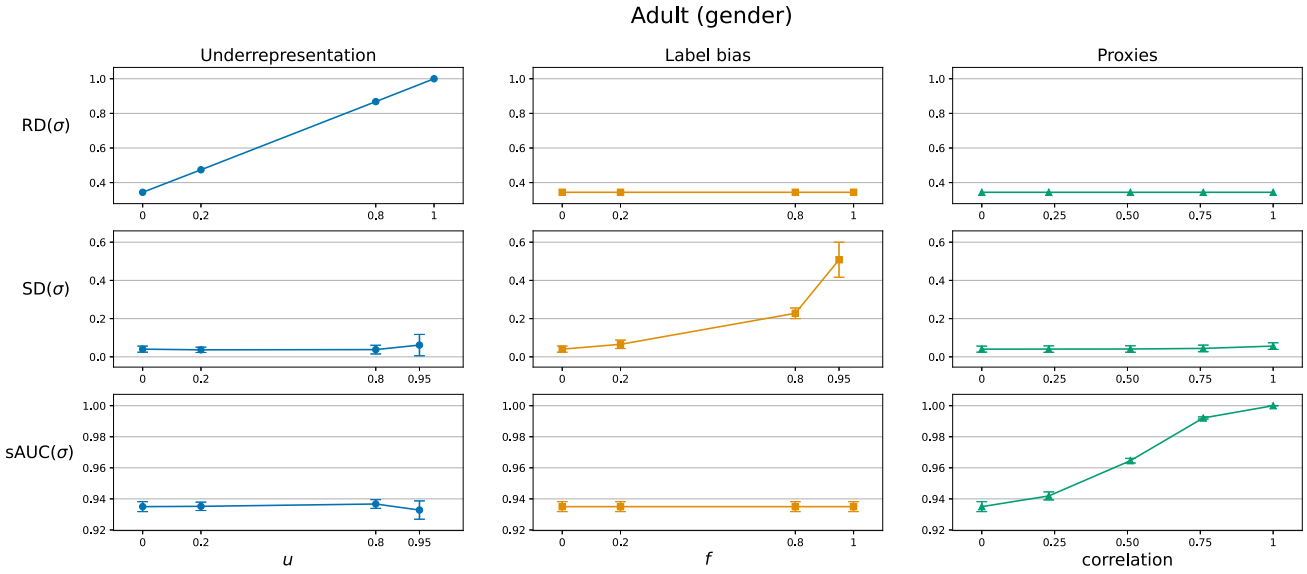

Fig. C.6. Results of the bias detection methods on Adult (gender). The columns represent the three different scenarios while the rows represent the three bias detection methods.

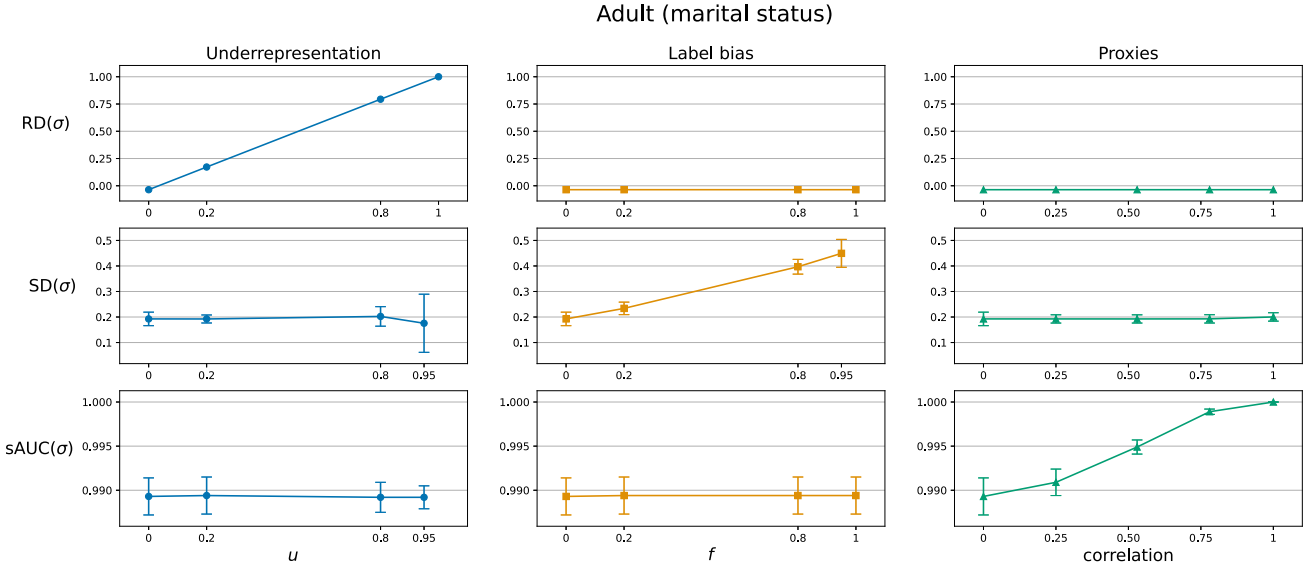

Fig. C.7. Results of the bias detection methods on Adult (marital status). The columns represent the three different scenarios while the rows represent the three bias detection methods.

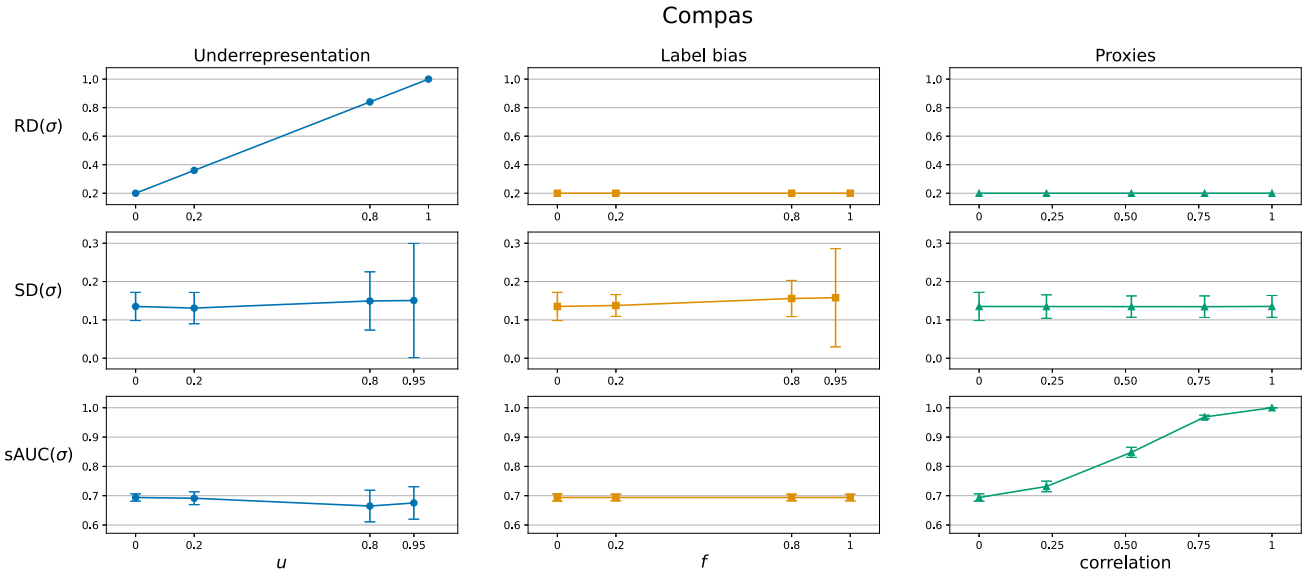

Fig. C.8. Results of the bias detection methods on Compas. The columns represent the three different scenarios while the rows represent the three bias detection methods.

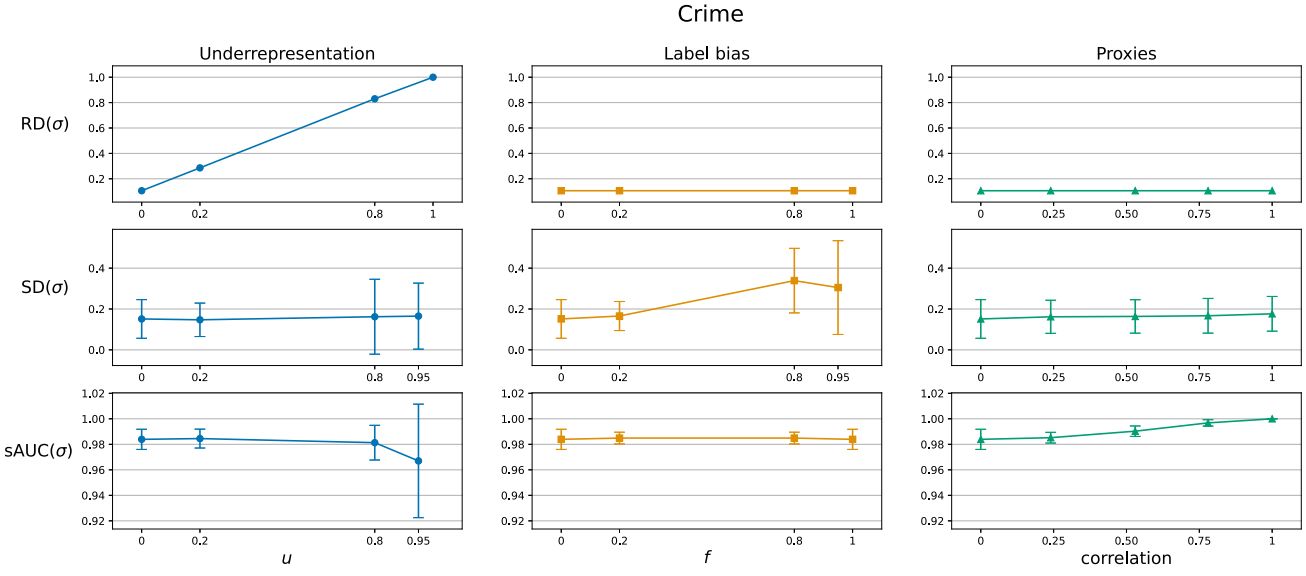

Fig. C.9. Results of the bias detection methods on Crime. The columns represent the three different scenarios while the rows represent the three bias detection methods.

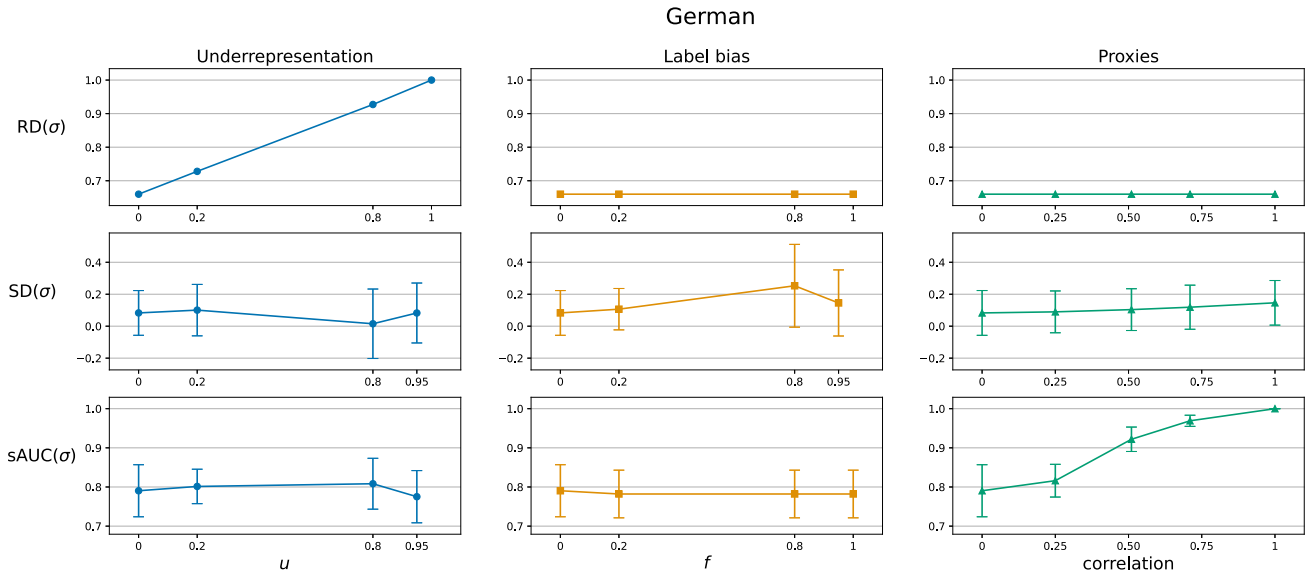

Fig. C.10. Results of the bias detection methods on German. The columns represent the three different scenarios while the rows represent the three bias detection methods.

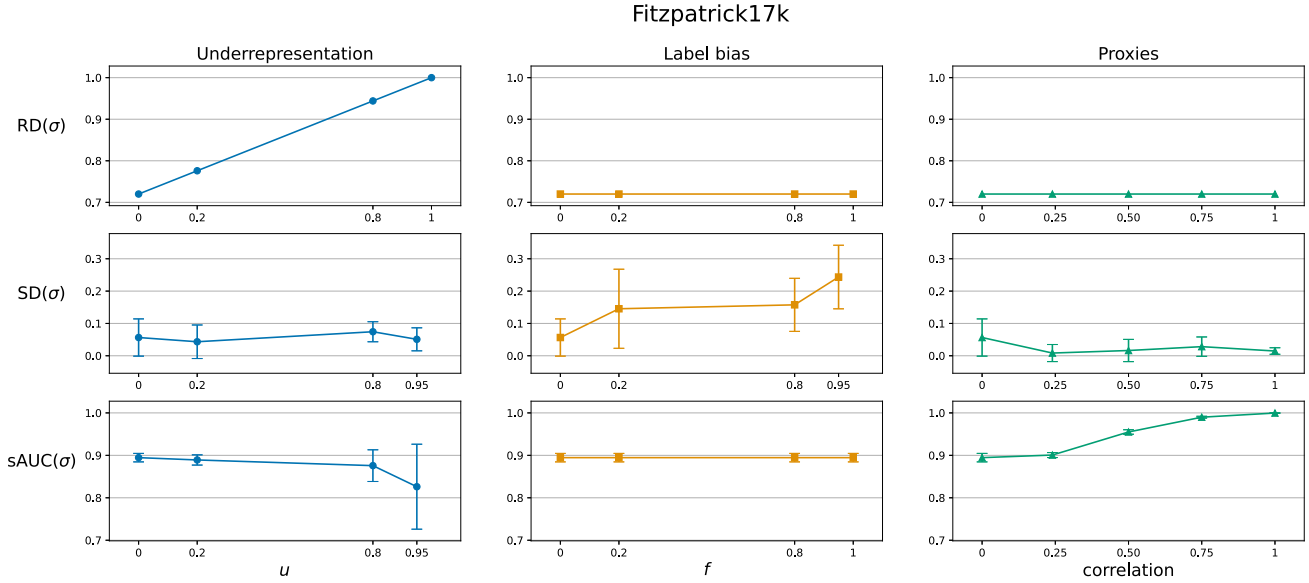

Fig. C.11. Results of the bias detection methods on Fitzpatrick17k. The columns represent the three different scenarios while the rows represent the three bias detection methods.

References

Álvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Lobo, P. R., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology*, *26*(2), 31. https://doi.org/10.1007/S10676-024-09746-W

Alves, G., Amblard, M., Bernier, F., Couceiro, M., & Napoli, A. (2021). Reducing unintended bias of ML models on tabular and textual data. In *8th IEEE International conference on data science and advanced analytics, DSAA 2021, Porto, Portugal, October 6–9, 2021* (pp. 1–10). IEEE. https://doi.org/10.1109/DSAA53316.2021.9564112

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/S40593-021-00285-9

Balunovic, M., Ruoss, A., & Vechev, M. T. (2022). Fair normalizing flows. In *The tenth international conference on learning representations, ICLR 2022, virtual event, April 25–29, 2022*. OpenReview.net. https://openreview.net/forum?id=BrFIKuxrZE.

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., & Venkatasubramanian, S. (2021). It's compaslicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In J. Vanschoren, & S. Yeung (Eds.), *Proceedings of the neural information processing systems track on datasets and benchmarks 1, NeurIPS datasets and benchmarks 2021, December 2021, virtual*. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/92cc227532d17e56e07902b254dfad10-Abstract-round1.html.

Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *,California Law Review 104*, 671.

Baumann, J., Castelnovo, A., Crupi, R., Inverardi, N., & Regoli, D. (2023). Bias on demand: A modelling framework that generates synthetic data with bias. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, FAccT 2023, Chicago, IL, USA, June 12–15, 2023* (pp. 1002–1013). ACM. https://doi.org/10.1145/3593013.3594058

Blind Stairs (2024). Unbiased interviews. https://blindstairs.com/en/interviews

Bogen, M. (2024). Navigating demographic measurement for fairness and equity. https://cdt.org/wp-content/uploads/2024/05/2024-04-29-AI-Gov-Lab-Demographic-Data-report-final.pdf.

Borgesius, F. Z., Baranowska, N., Hacker, P., & Fabris, A. (2024). Non-discrimination law in Europe: A primer. introducing European non-discrimination law to non-lawyers. arXiv preprint arXiv:2404.08519.

Brzezinski, D., Stachowiak, J., Stefanowski, J., Szczech, I., Susmaga, R., Aksenyuk, S., Ivashka, U., & Yasinskyi, O. (2024). Properties of fairness measures in the context of varying class imbalance and protected group ratios. *ACM Transactions on Knowledge Discovery from Data*, *18*, 170.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler, & C. Wilson (Eds.), *Conference on fairness, accountability and transparency, FAT 2018, 23–24 February 2018, New York, USA* (pp. 77–91). PMLR (*vol. 81*). Proceedings of Machine Learning Research. http://proceedings.mlr.press/v81/buolamwini18a.html.

Chen, Y., Giudici, P., Liu, K., & Raffinetti, E. (2024). Measuring fairness in credit ratings. *Expert Systems with Applications*, *258*, 125184. https://doi.org/10.1016/J.ESWA.2024.125184

Cobham, A. (2020). The uncounted. Polity.

Cooper, A. F., Lee, K., Choksi, M. Z., Barocas, S., De Sa, C., Grimmelmann, J., Kleinberg, J. M., Siddhartha, S., & Zhang, B. (2024). Arbitrariness and social prediction: The confounding role of variance in fair classification. In M. J. Wooldridge, J. G. Dy, & S. Natarajan (Eds.), *Thirty-eighth AAAI conference on artificial intelligence, AAAI 2024, thirty-sixth conference on innovative applications of artificial intelligence, IAAI 2024, fourteenth symposium on educational advances in artificial intelligence, EAAI 2014, February 20–27, 2024, Vancouver, Canada* (pp. 22004–22012). AAAI Press. https://doi.org/10.1609/AAAI.V38I20.30203

Cornacchia, G., Anelli, V. W., Biancofiore, G. M., Narducci, F., Pomo, C., Ragone, A., & Sciascio, E. D. (2023). Auditing fairness under unawareness through counterfactual reasoning. *Information Processing & Management*, *60*(2), 103224. https://doi.org/10.1016/J.IPM.2022.103224

Cruz, A. F., & Hardt, M. (2024). Unprocessing seven years of algorithmic fairness. In *The twelfth international conference on learning representations, ICLR 2024, Vienna, Austria, 7–11, 2024*. OpenReview.net. https://openreview.net/forum?id=jr03SfWsBS.

Deck, L., Müller, J.-L., Braun, C., Zipperling, D., & Kühl, N. (2024). Implications of the AI act for non-discrimination law and algorithmic fairness. arXiv preprint arXiv:2403.20089 .

Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, *34*, 496.

Dominguez-Catena, I., Paternain, D., & Galar, M. (2024). Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5209–5226. https://doi.org/10.1109/TPAMI.2024.3361979

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., & Pontil, M. (2018). Empirical risk minimization under fairness constraints. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp. 2796–2806). https://proceedings.neurips.cc/paper/2018/hash/83cdcec08fbf90370fcf53bdd56604ff-Abstract.html.

Drukker, K., Chen, W., Gichoya, J., Gruszauskas, N., Kalpathy-Cramer, J., Koyejo, S., Myers, K., Sá, R. C., Sahiner, B., Whitney, H. et al. (2023). Toward fairness in artificial intelligence for medical image analysis: Identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging*, *10*(6), 061104.

Du, M., Yang, F., Zou, N., & Hu, X. (2021). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, *36*(4), 25–34. https://doi.org/10.1109/MIS.2020.3000681

Edwards, H., & Storkey, A. J. (2016). Censoring representations with an adversary. In Y. Bengio, & Y. LeCun (Eds.), *4th International conference on learning representations, ICLR 2016, San Juan, Puerto Rico, 2–4, 2016, conference track proceedings*. http://arxiv.org/abs/1511.05897.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J. M., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115–118. https://api.semanticscholar.org/CorpusID:3767412.

Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F., & Biega, A. J. (2024). Fairness and bias in algorithmic hiring. *ACM Transactions on Intelligent Systems and Technology*, https://doi.org/10.1145/3696457.

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery*, *36*(6), 2074–2152. https://doi.org/10.1007/S10618-022-00854-Z

Fabris, A., Silvello, G., Susto, G. A., & Biega, A. J. (2023). Pairwise fairness in ranking as a dissatisfaction measure. In T. Chua, H. W. Lauw, L. Si, E. Terzi, & P. Tsaparas (Eds.), *Proceedings of the sixteenth ACM international conference on web search and data mining, WSDM 2023, Singapore, 27 February 2023 - 3 March 2023* (pp. 931–939). ACM. https://doi.org/10.1145/3539597.3570459

Fajri, R. M., Saxena, A., Pei, Y., & Pechenizkiy, M. (2024). FAL-CUR: Fair active learning using uncertainty and representativeness on fair clustering. *Expert Systems with Applications*, *242*, 122842. https://doi.org/10.1016/J.ESWA.2023.122842

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, & G. Williams (Eds.), *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, NSW, Australia, August 10–13, 2015* (pp. 259–268). ACM. https://doi.org/10.1145/2783258.2783311

Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology, 124 6*, 869–871. https://api.semanticscholar.org/CorpusID:29991932.

Fragkathoulas, C., Papanikou, V., Karidi, D. P., & Pitoura, E. (2024). On explaining unfairness: An overview. In *40th International conference on data engineering, ICDE 2024 - workshops, Utrecht, Netherlands, May 13–16, 2024* (pp. 226–236). IEEE. https://doi.org/10.1109/ICDEW61823.2024.00035

Gajane, P., & Pechenizkiy, M. (2018). On formalizing fairness in prediction with machine learning. arXiv preprint:1710.03184.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, *178*(11), 1544–1547.

Gillis, T. B., Meursault, V., & Ustun, B. (2024). Operationalizing the search for less discriminatory alternatives in fair lending. In *The 2024 ACM conference on fairness, accountability, and transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3–6, 2024* (pp. 377–387). ACM. https://doi.org/10.1145/3630106.3658912

Glazko, K. S., Mohammed, Y., Kosa, B., Potluri, V., & Mankoff, J. (2024). Identifying and improving disability bias in GPT-based resume screening. In *The 2024 ACM conference on fairness, accountability, and transparency, FAccT 2024, Rio De Janeiro, Brazil, June 3–6, 2024* (pp. 687–700). ACM. https://doi.org/10.1145/3630106.3658933

Golpayegani, D., Hupont, I., Panigutti, C., Pandit, H. J., Schade, S., O'Sullivan, D., & Lewis, D. (2024). AI Cards: Towards an applied framework for machine-readable AI and risk documentation inspired by the EU AI act. In M. Jensen, C. Lauradoux, & K. Rannenberg (Eds.), *Privacy technologies and policy - 12th annual privacy forum, APF 2024, Karlstad, Sweden, September 4–5, 2024, proceedings* (pp. 48–72). Springer (*vol. 14831*). Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-68024-3_3

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1820–1828).

Guerdan, L., Coston, A., Wu, Z. S., & Holstein, K. (2023). Ground(less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, FAccT 2023, Chicago, IL, USA, June 12–15, 2023* (pp. 688–704). ACM. https://doi.org/10.1145/3593013.3594036

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29: Annual conference on neural information processing systems 2016, December 5–10, 2016, Barcelona, Spain* (pp. 3315–3323). https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

HireVue (2022). Explainability statement. https://hirevue-api.dev-directory.com/wp-content/uploads/2022/04/HV_AI_Short-Form_Explainability_1pager.pdf.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2020). The dataset nutrition label. *Data Protection and Privacy*, *12*(12), 1.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible*

Computing, *1*(2). https://doi.org/10.1145/3631326. https://doi.org/10.1145/3631326

Hu, J., Zeng, H., Li, H., Niu, C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, & P. J. Shenoy (Eds.), *Proceedings of the 16th international conference on world wide web, WWW 2007, Banff, Alberta, Canada, May 8–12, 2007* (pp. 151–160). ACM. https://doi.org/10.1145/1242572.1242594

ISO (2021). Information technology — artificial intelligence (AI) — bias in AI systems and AI aided decision making. https://www.iso.org/standard/77607.html.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In M. C. Elish, W. Isaac, & R. S. Zemel (Eds.), *FAccT '21: 2021 ACM conference on fairness, accountability, and transparency, virtual event / Toronto, Canada, March 3–10, 2021* (pp. 375–385). ACM. https://doi.org/10.1145/3442188.3445901

cjadams, Borkan, D., inversion, Sorensen, J., Dixon, L., Vasserman, L., nithum. (2019). Jigsaw Unintended Bias in Toxicity Classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification Kaggle

Kallus, N., & Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. https://arxiv.org/abs/1902.05826.

Kamiran, F., & Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/S10115-011-0463-8

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In E. Simoudis, J. Han, & U. M. Fayyad (Eds.), *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96), Portland, Oregon, USA* (pp. 202–207). AAAI Press. http://www.aaai.org/Library/KDD/1996/kdd96-033.php.

Königstorfer, F., & Thalmann, S. (2022). AI documentation: A path to accountability. *Journal of Responsible Technology*, *11*, 100043.

Liu, Z., Qiu, R., Zeng, Z., Zhu, Y., Hamann, H., & Tong, H. (2024). Aim: Attributing, interpreting, mitigating data unfairness. *arXiv e-prints*, (pp. arXiv–2406).

Madras, D., Creager, E., Pitassi, T., & Zemel, R. S. (2018). Learning adversarially fair and transferable representations. In J. G. Dy, & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018* (pp. 3381–3390). PMLR (*vol. 80*). Proceedings of Machine Learning Research. http://proceedings.mlr.press/v80/madras18a.html.

Mecati, M., Torchiano, M., Vetrò, A., & De Martin, J. C. (2023). Measuring imbalance on intersectional protected attributes and on target variable to forecast unfair classifications. *IEEE Access*, *11*, 26996–27011. https://doi.org/10.1109/ACCESS.2023.3252370

Mecati, M., Vetrò, A., & Torchiano, M. (2022). Detecting risk of biased output with balance measures. *ACM Journal of Data and Information Quality*, *14*(4), 25:1–25:7. https://doi.org/10.1145/3530787

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115:1–115:35. https://doi.org/10.1145/3457607

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Parliament, E. (2024). Artificial intelligence act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In Y. Li, B. Liu, & S. Sarawagi (Eds.), *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, Las Vegas, Nevada, USA, August 24–27, 2008* (pp. 560–568). ACM. https://doi.org/10.1145/1401890.1401959

Perez, C. C. (2019). Invisible women: Data bias in a world designed for men. Abrams.

Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. In *FAccT '22: 2022 ACM conference on fairness, accountability, and transparency, Seoul, Republic of Korea, June 21–24, 2022* (pp. 1776–1826). ACM. https://doi.org/10.1145/3531146.3533231

Rondina, M., Vetrò, A., & De Martin, J. C. (2023). Completeness of datasets documentation on ML/AI repositories: An empirical investigation. In N. Moniz, Z. Vale, J. Cascalho, C. Silva, & R. Sebastião (Eds.), *Progress in artificial intelligence - 22nd EPIA conference on artificial intelligence, EPIA 2023, Faial Island, Azores, September 5–8, 2023, proceedings, Part I* (pp. 79–91). Springer (*vol. 14115*). Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-49008-8_7

Ruoss, A., Balunovic, M., Fischer, M., & Vechev, M. T. (2020). Learning certified individually fair representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/55d491cf951b1b920900684d71419282-Abstract.html.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., & Aroyo, L. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. M. Drucker (Eds.), *CHI '21: CHI Conference on human factors in computing systems, virtual event / Yokohama, Japan, May 8–13, 2021* (pp. 39:1–39:15). ACM. https://doi.org/10.1145/3411764.3445518

Santamaría, L., & Mihaljevic, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156. https://doi.org/10.7717/PEERJ-CS.156

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. US Department of Commerce, National Institute of Standards and Technology.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y., & Ghassemi, M. (2020). Chexclusion: Fairness gaps in deep chest x-ray classifiers. *Biocomputing 2021* 232–243.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine learning for the developing world*.

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *EAAMO 2021: ACM Conference on equity and access in algorithms, mechanisms, and optimization, virtual event, USA, October 5–9, 2021* (pp. 17:1–17:9). ACM. https://doi.org/10.1145/3465416.3483305

Vardasbi, A., de Rijke, M., Diaz, F., & Dehghani, M. (2024). The impact of group membership bias on the quality and fairness of exposure in ranking. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval (SIGIR '24),*. ACM. https://doi.org/10.1145/3626772.3657752

Vetrò, A., Torchiano, M., & Mecati, M. (2021). A data quality approach to the identification of discrimination risk in automated decision making systems. *Government Information Quarterly*, *38*(4), 101619. https://doi.org/10.1016/J.GIQ.2021.101619

Wang, J., Liu, Y., & Levy, C. C. (2021). Fair classification with group-dependent label noise. In M. C. Elish, W. Isaac, & R. S. Zemel (Eds.), *Facct '21: 2021 ACM conference on fairness, accountability, and transparency, virtual event / Toronto, Canada, March 3–10, 2021* (pp. 526–536). ACM. https://doi.org/10.1145/3442188.3445915

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097–2106).

Weng, N., Bigdeli, S., Petersen, E., & Feragen, A. (2023). Are sex-based physiological differences the cause of gender bias for chest x-ray diagnosis? In *Workshop on clinical image-based procedures* (pp. 142–152). Springer.

Yin, T., Raab, R., Liu, M., & Liu, Y. (2023). Long-term fairness with unknown dynamics. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems 36: Annual conference on neural information processing systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*. http://papers.nips.cc/paper_files/paper/2023/hash/acf4a08f67724e9d2de34099f57a9c25-Abstract-Conference.html.

Zhang, G., Cheng, D., Yuan, G., & Zhang, S. (2024). Learning fair representations via rebalancing graph structure. *Information Processing & Management*, *61*(1), 103570. https://doi.org/10.1016/J.IPM.2023.103570

Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089. https://doi.org/10.1007/S10618-017-0506-1