Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey

Max Hort, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Mark Harman

Abstract—This paper provides a comprehensive survey of bias mitigation methods for achieving fairness in Machine Learning (ML) models. We collect a total of 341 publications concerning bias mitigation for ML classifiers. These methods can be distinguished based on their intervention procedure (i.e., pre-processing, in-processing, post-processing) and the technology they apply. We investigate how existing bias mitigation methods are evaluated in the literature. In particular, we consider datasets, metrics and benchmarking. Based on the gathered insights (e.g., What is the most popular fairness metric? How many datasets are used for evaluating bias mitigation methods?). We hope to support practitioners in making informed choices when developing and evaluating new bias mitigation methods.

Index Terms—fairness, bias mitigation, machine learning

1 Introduction

Machine Learning (ML) has been increasingly popular in recent years, both in the diversity and importance of applications [1]. ML is used in a variety of critical decision-making applications including justice risk assessments [2], [3] and job recommendations [4].

While ML systems have the advantage to relieve humans from tedious tasks and are able to perform complex calculations at a higher speed [5], they are only as good as the data on which they are trained [6]. ML algorithms, which are never designed to intentionally incorporate bias, run the risk of replicating or even amplifying bias present in real-world data [6], [7], [8]. This may cause unfair treatment in which some individuals or groups of people are *privileged* (i.e., receive a favourable treatment) and others are *unprivileged* (i.e., receive an unfavourable treatment). In this context, a fair treatment of individuals constitutes that decisions are made independent of sensitive attributes such as gender or race, such that individuals are treated based on merit [9], [10], [11]. For example, one can aim for an equal probability of population groups to receive a positive treatment, or an equal treatment of individuals that only differ in sensitive attributes.

Human bias has been transferred to various real-word systems relying on ML. There are many examples of this in the literature. For instance, bias has been found in advertisement and recruitment processes [4], [12], affecting university admissions [13] and human rights [11]. Not only is such a biased behaviour undesired, but it can fall under regulatory control and risk the violation of anti-discrimination laws [7], [14], [15], as sensitive attributes such as age, disability, gender identity, race are protected by US law in the Fair Housing Act and Equal Credit Opportunity Act [16].

Another example for a biased treatment of population groups can be found in the **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) software, used by courts in US to determine the risks of an individual to reoffend. These scores are used to motivate decisions on whether and when defendants are to be set free, in different stages of the justice system. Problematically, this software falsely labelled non-white defendants with higher risk scores than white defendants [2].

To reduce the degree of bias that such systems exhibit, practitioners use three types of bias mitigation methods [17]:

- **Pre-processing:** bias mitigation in the training data, to prevent it from reaching ML models;
- **In-processing:** bias mitigation while training ML models;
- **Post-processing:** bias mitigation on trained ML models.

There has been a growing interest in fairness research, including definitions, measurements, and improvements of ML models [1], [5], [18], [19], [20]. In particular, a variety of recent work addresses the mitigation of bias in binary classification models: given a collection of observations (training data) are labelled with a binary label (testing data) [21].

Despite the large amount of existing bias mitigation methods and surveys on fairness research, as Pessach and Shmueli [5] pointed out, there remain open challenges that practitioners face when designing new bias mitigation methods: "It is not clear how newly proposed mechanisms should be evaluated, and in particular which measures should be considered? which datasets should be used? and which mechanisms should be used for comparison?" [5]

To combat this challenge, we set out to perform a comprehensive survey of existing research on bias mitigation for ML models. We analyse 341 publications to identify

- *Max Hort is with the Department of Computer Science, University College London, London, United Kingdom. E-mail: max.hort.19@ucl.ac.uk*
- *Zhenpeng Chen is with the Department of Computer Science, University College London, London, United Kingdom. E-mail: zp.chen@ucl.ac.uk*
- *Jie M. Zhang is with the Department of Informatics, King's College London, London, United Kingdom. E-mail: jie.zhang@kcl.ac.uk*
- *Federica Sarro is with the Department of Computer Science, University College London, London, United Kingdom. E-mail: f.sarro@ucl.ac.uk*
- *Mark Harman is with the Department of Computer Science, University College London, London, United Kingdom. E-mail: mark.harman@ucl.ac.uk*

Manuscript received xxx; revised xxx.

practices applied in fairness research when creating bias mitigation methods. In particular, we consider the datasets to which bias mitigation methods are applied, the metrics used to determine the degree of bias, and the approaches used for benchmarking the effectiveness of bias mitigation methods. By doing so, we allow practitioners to focus their effort on creating bias mitigation methods rather than requiring a lot of time to determine their experimental setup (e.g., which datasets to test on, which benchmark to consider).

To the best of our knowledge, this is the first survey to systematically and comprehensively cover bias mitigation methods and their evaluation. To summarize, the contribution of this survey are:

1) we provide a comprehensive overview of the research on bias mitigation methods for ML classifiers;
2) we introduce the experimental design details for evaluating existing bias mitigation methods;
3) we identify challenges and opportunities for future research on bias mitigation methods.
4) we make the collected paper repository public, to allow for future replication and manual investigation of our results [22].

The rest of this paper is structured as follows. Section 2 presents an overview of related surveys. The search methodology is described in Section 3. Sections 4-7 describe research on bias mitigation methods. Challenges that the field of fairness research and bias mitigation methods face are discussed in Section 8. Section 9 concludes this survey.

2 Related Surveys

In this section, we provide an overview on existing surveys in the fairness literature and their contents. This allows us to identify the knowledge gap filled by our survey.

Mehrabi et al. [11] and Pessach and Shmueli [5] provided an overview of bias and discrimination types, fairness definitions and metrics, bias mitigation methods, and existing datasets. For example, Pessach and Shmueli [5], [23] listed the datasets and metrics used by 27 bias mitigation methods. A similar focus has been pursued by Dunkelau and Leuschel [18], who provided an extensive overview on fairness notions, available frameworks, and bias mitigation methods for classification problems. They moreover provided a classification of approaches for each type (i.e., pre-, in-, and post-processing). The most exhaustive categorization of bias mitigation methods, to date, has been conducted by Caton and Haas [24], who also presented fairness metrics and fairness platforms.

A detailed collection of prominent fairness definitions for classification problems is provided by Verma and Rubin [21]. Similarly, Žliobaite [25] surveyed measures for indirect discrimination for ML.

In addition to the surveys on fairness metrics, Le Quy et al. [26] provided a survey with 15 frequently used datasets in fairness research. For each dataset, they described the available features and their relationships with sensitive attributes.

Other surveys are concerned with fairness and consider the following perspectives: learning-based sequential decision algorithms [27], criminal justice [3], graph representations [28], ML testing [29], Software Engineering [30], [31], or Natural Language Processing [32], [33].

While previous surveys focus on ML classification, and some mention bias mitigation methods, none has yet systematically covered the evaluation bias mitigation methods (e.g., how are methods benchmarked, what dataset are used). The surveys related closest to our focus are provided by Dunkelau and Leuschel [18], and Pessach and Shmueli [5], [23].

Dunkelau and Leuschel [18] provided an overview of bias mitigation methods, with a focus on their implementation and underlying algorithms. However, further evaluation details of these methods, such as dataset and metric usage, were not addressed. While Pessach and Shmueli [5], [23] listed the datasets and metrics used by 27 bias mitigation methods, they do not provide actionable insights to support developers. In addition to combining aspects of both surveys (i.e., extensive collection of bias mitigation methods like Dunkelau and Leuschel [18], and information on datasets and metrics similar to Pessach and Shmueli [5]), we aim to analyze the findings of a comprehensive literature search to devise recommendations.

3 Survey Methodology

The purpose of this survey is to gather and categorize research work, that mitigates bias in ML models. Given that the existing literature focuses on classification for tabular data, this survey also focuses on bias mitigation methods for such classification tasks.

3.1 Search Methodology

This section outlines our search procedure. We start with a preliminary search, followed by a repository search and snowballing.

Preliminary Search. Prior to systematically searching online repositories, we conduct a preliminary search. The goal of the preliminary search is to gain a deeper understanding of the field and assess whether there is a sufficient number of publications to allow for subsequent analysis. In particular, we collect bias mitigation publications from four existing surveys (see Section 2):

- Mehrabi et al. [11] : 24 bias mitigation methods;
- Pessach and Shmueli [5], [23]: 30 bias mitigation methods;
- Dunkelau and Leuschel [18]: 40 bias mitigation methods;
- Caton and Haas [24]: 70 bias mitigation methods.

In total, we collect 100 unique bias mitigation methods from these four surveys.

Repository Search. After the preliminary search, we conduct a search of six established online repositories (IEEE, ACM, ScienceDirect, Scopus, arXiv, and Google Scholar).

The search procedure is guided by two groups of keywords:

- Domain: machine learning, deep learning, artificial intelligence;

TABLE 1: Publications found at each stage of the search procedure.

Stage	Publications
Preliminary search	100
Repository search Oct'21	75
Repository search Jul'22	56
Snowballing	78
Author feedback	32
Total	341

- Bias Mitigation: fairness-aware, discrimination-aware, bias mitigation, debias*, unbias*;

In this context, *Domain* keywords ensure that the bias discussed in the publication affects machine learning systems. *Bias Mitigation* ensures that the publication addresses bias reduction via the use of bias mitigation methods. For the six repositories, we collected publications that contain at least one *Domain* and one *Bias mitigation* keyword (i.e., we check each possible combination of keywords for the two categories).

Selection To ensure that the publications included in this survey are relevant to the context of bias mitigation for ML models, we consider the following **inclusion criteria**: 1) describe human biases; 2) address classification problems; 3) use tabular data (e.g., do not make decisions based on images or text alone).

To ensure that irrelevant publications are excluded from the search results, we manually check publications in three filtration stages [34]:

1) **Title:** Publications with irrelevant titles to the survey are excluded;
2) **Abstract:** The abstract of every publication is checked. Publications that show to be irrelevant to the survey at this step are excluded (e.g. not about ML, do not apply debiasing);
3) **Body:** For publications that passed the previous two steps, we check the entire publication to determine whether they satisfy the inclusion criteria. If not, they are excluded.

Snowballing After conducting the repository search, we apply backward snowballing (i.e., finding new publications that are cited by publications we already selected) for each publication retained after the "Body" stage [35]. This snowballing step is repeated for every new publication found. The goal of snowballing is to find missing related work with regards to the collected publications. This is in particular useful if undiscovered bias mitigation methods are used for benchmarking.

3.2 Selected Publications

In total, we gathered 341 publications over the different stages of our search procedure. Table 2 summarises the results of two repository searches. The first search was conducted from the 7th of October to 10th of October 2021, and the second search was conducted on the 21st of July 2022. The purpose of the second search is to collect publications from the year 2022 (i.e., we filtered search results for the publication year 2022). In October 2021, Google Scholar provided $8,738$ publications that were in line with the search keywords. We restricted our search to the first $1,000$ entries as prioritised by Google Scholar based on relevance. Similarly, the second search yielded $1,995$ results and we focused on the first $1,000$ publications.

To ensure that our survey is comprehensive and accurate, we contacted the corresponding authors of the 309 publications collected via the preliminary search, the two repository searches and snowballing. We asked them to check whether our description about their work is correct. Based on their feedback, we included additional 31 publications. The amount of publications found for each step of the search is listed in Table 1.

TABLE 2: Results of the repository search. For each of the six search repositories, we show the number of publications retained after each filtration stage, where the "Body" column shows the number of publications included in this survey.

Repository	Initial	Title	Abstract	Body
ACM	118	26	16	13
ScienceDirect	166	9	5	3
IEEE	401	18	9	9
arXiv	650	69	48	38
Scopus	1063	44	28	21
Google Scholar	8738	119	90	77

Search results October'21.

Repository	Initial	Title	Abstract	Body
ACM	468	17	14	8
ScienceDirect	88	6	3	2
IEEE	90	8	1	1
arXiv	465	42	23	17
Scopus	356	13	9	5
Google Scholar	1995	62	51	35

Search results July'22.

4 Algorithms

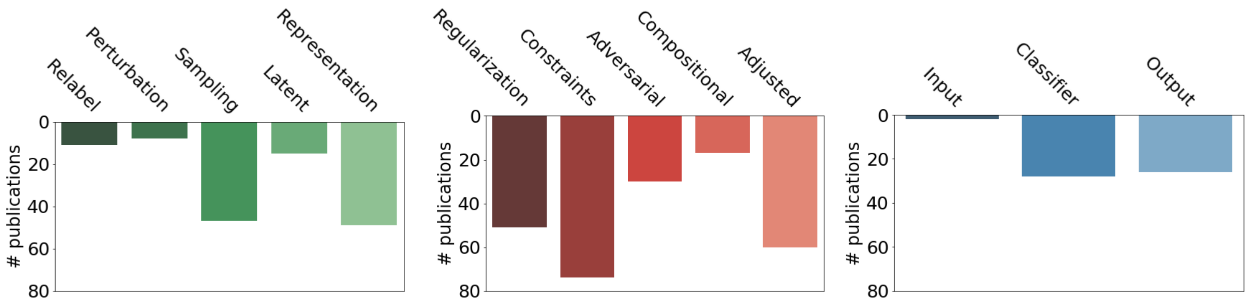
In this section, we present the bias mitigation methods found in our literature search. We distinguished bias mitigation methods based on their type (i.e., in which stage of the ML process are they applied): pre-processing (Section 4.1), in-processing (Section 4.2) and post-processing (Section 4.3) methods [17] . Moreover, we organize methods in categories (i.e., the bias mitigation approach). For this, we follow taxonomies devised by Dunkelau and Leuschel [18], as well as Caton and Haas [24]. Figure 1 illustrates the 13 categories we use.

A single publication may reside in multiple categories, for example if their approach applies pre-processing before adapting the training procedure during an in-processing stage. This is the case for 70 publications, for which we provide more information in Section 4.4.

Among the 341 publications, 123 used pre-processing (Section 4.1), 212 used in-processing (Section 4.2) and 56 used post-processing methods (Section 4.3).

4.1 Pre-processing Bias Mitigation Methods

In this section, we present bias mitigation methods that combat bias by applying changes to the training data. Table 3



(a) Pre-processing. (b) In-processing. (c) Post-processing.

Fig. 1: Categorization of bias mitigation methods. Categories are grouped based on their type (i.e., pre-processing, in-processing, post-processing) and the number of publications of each category is shown.

and Table 4 list the 123 publications we found, according to the type of pre-processing method used.

4.1.1 Relabelling and Perturbation

This section presents bias mitigation methods that apply changes to the values of the training data. Changes have been applied to the ground truth labels (relabelling) or the remaining features (perturbation).

A popular approach for relabelling the dataset is "massaging", proposed by Kamiran and Calders [36] in 2009. In the first stage, "massaging" uses a ranker to determine the best candidates for relabelling. In particular, instances close to the decision boundary are selected, to minimize the negative impact of relabelling on accuracy. Afterwards, an equal amount of instances with positive and negative labels are typically selected, according to their rank. For selected instances, their labels are switched.

Massaging has later been extended by Kamiran and Calders [41], and Calders et al. [37]. Moreover, Žliobaite et al. [39] created a related method called "local massaging". "Massaging" has also been applied by other work [42], [43].

Another relabelling approach was proposed by Loung et al. [38], who relabelled instances based on their k-nearest neighbours, such that similar individuals receive similar labels.

Feldman et al. [47] used perturbation to modify non-protected attributes, such that their values for privileged and unprivileged groups are comparable. In particular, the values are adjusted to bring their distributions closer together while preserving the respective ranks within a group (e.g., the highest values of attribute a for the privileged group remains highest after perturbation). Lum and Johndrow [48], [51] used conditional models for perturbation, which allowed for modification of multiple variables (continuous or discrete). Li et al. [52] proposed an iterative approach for perturbation. At each step, the most bias-prone attribute is selected and transformed, until the degree of bias exhibited by a classification model is below a specified threshold.

Other than perturbing the underlying data for all groups to move them closer [47], [48], [51], Wang et al. [49], [50] considered only the unprivileged group for perturbation seeking to resolve disparity by improving the performance of the unprivileged group. Hajian et al. [40] applied both relabeling and perturbation (i.e., changes to the sensitive attribute).

4.1.2 Sampling

Sampling methods change the training data by changing the distribution of samples (e.g., adding, removing samples) or adapting their impact on training. Similarly, the impact of training data instances can be achieved by reweighing their importance [37], [41], [43], [72], [73], [82], [85], [87], [91], [94], [95].

Reweighing was first introduced by Calders et al. [37]. Each instance receives a weight according to its label and protected attribute (e.g., instances in the unprivileged group and positive label receive a higher weight as this is less likely). In the training process of classification models, a higher instance weight causes higher losses when misclassified. Weighted instances are sampled with replacement according to their weights. If the classification model is able to process weighted instances, the dataset can be used for training without resampling [41].

Jiang and Nachum [68] and Krasanakits et al. [56] used reweighing to combat biased labels in the original training data.

Instead of assigning equal weights to data instances of the same population subgroup, Li et al. [91] assigned individual weights to instances of the training data.

Other sampling strategies include the removal of data points (downsampling) [60], [67], [75], [78], [79], [83], [84], [90], [93] or the addition of new data points (upsampling). Popular methods for upsamplig are oversampling for duplicating instances of the minority group [59], [61], [70], [77] and the use of SMOTE [159]. SMOTE does not duplicate instances but generates synthetic ones in the neighborhood of the minority group [59], [61], [70], [71], [74], [80], [86], [89], [92].

To sample datapoints, uniform [41] and preferential [39], [41], [54], [61], [69] strategies have been followed, where preferential sampling changes the distribution of instances close to the decision boundary.

Xu et al. [57], [62], [63] used a generative approach to generate discrimination-free data for training [65], [81], [88]. Zhang et al. [55] used causal networks to create a new dataset. The initial dataset is used to create a causal network, which is

TABLE 3: Publications on Pre-processing bias mitigation methods.

Category	Authors [Ref]	Year	Venue
Relabel	Kamiran and Calders [36]	2009	ICCCC
	Calders et al. [37]	2009	ICDMW
	Loung et al. [38]	2011	KDD
	Žliobaite et al. [39]	2011	ICDM
	Hajian et al. [40]	2012	IEEE Trans Knowl Data Eng
	Kamiran and Calders [41]	2012	KAIS
	Zhang et al. [42]	2018	IJCAI
	Iosifidis et al. [43]	2019	DEXA
	Sun et al. [44]	2022	EuroS&P
	Seker et al. [45]	2022	Stud. Health Technol. Inform
	Alabdulmohsin et al. [46]	2022	arXiv
Perturbation	Hajian et al. [40]	2012	IEEE Trans Knowl Data Eng
	Feldman et al. [47]	2015	KDD
	Lum and Johndrow [48]	2016	arXiv
	Wang et al. [49]	2018	NeurIPS
	Wang et al. [50]	2019	ICML
	Johndrow and Lum [51]	2019	Ann Appl Stat
	Li et al. [52]	2022	SSRN
	Li et al. [53]	2022	ICSE
Sampling	Calders et al. [37]	2009	ICDMW
	Kamiran and Calders [54]	2010	BNAIC
	Žliobaite et al. [39]	2011	ICDM
	Kamiran and Calders [41]	2012	KAIS
	Zhang et al. [55]	2017	IJCAI
	Krasanakits et al. [56]	2018	TheWebConf
	Xu et al. [57]	2018	Big Data
	Chen et al. [58]	2018	NeurIPS
	Iosifidis and Ntoutsi [59]	2018	report
	Salimi et al. [60]	2019	MOD
	Iosifidis et al. [43]	2019	DEXA
	Zelaya et al. [61]	2019	KDD
	Xu et al. [62]	2019	IJCAI
	Xu et al. [63]	2019	Big Data
	Iosifidis et al. [64]	2019	Big Data
	Abusitta et al. [65]	2019	arXiv
	Sharma et al. [66]	2020	AIES
	Chakraborty et al. [67]	2020	FSE
	Jiang and Nachum [68]	2020	AISTATS
	Hu et al. [69]	2020	DS
	Morano [70]	2020	Thesis
	Yan et al. [71]	2020	CIKM
	Celis et al. [72]	2020	ICML
	Abay et al. [73]	2020	arXiv
	Salazar et al. [74]	2021	IEEE Access
	Zhang et al. [75]	2021	PAKDD
	Chuang and Mroueh [76]	2021	ICLR
	Amend and Spurlock [77]	2021	JCSC
	Verma et al. [78]	2021	arXiv
	Cruz et al. [79]	2021	ICDM
	Chakraborty et al. [80]	2021	FSE
	Jang et al. [81]	2021	AAAI
	Du and Wu [82]	2021	CIKM
	Roh et al. [83]	2021	NeurIPS
	Iofinova et al. [84]	2021	arXiv
	Yu [85]	2021	arXiv
	Singh et al. [86]	2021	Mach. learn. knowl. Extr.
	Sun et al. [44]	2022	EuroS&P
	Pentyala et al. [87]	2022	arXiv
	Rajabi et al. [88]	2022	Mach. learn. knowl. Extr.
	Dablain et al. [89]	2022	arXiv
	Chen et al. [90]	2022	FSE
	Li et al. [91]	2022	PMLR
	Chakraborty et al. [92]	2022	FairWARE
	Wang et al. [93]	2022	ICML
	Almuzaini et al. [94]	2022	FAccT
	Chai and Wang [95]	2022	ICML
Latent	Calders and Verwer [96]	2010	Data Min. Knowl. Discov
	Kilbertus et al. [97]	2017	NeurIPS
	Gupta et al. [98]	2018	arXiv
	Madras et al. [99]	2019	FAccT
	Oneto et al. [100]	2019	AIES
	Wei et al. [101]	2020	PMLR
	Kehrenberg et al. [102]	2020	Front. Artif. Intell.
	Grari et al. [103]	2021	arXiv
	Chen et al. [104]	2022	arXiv
	Liang et al. [105]	2022	arXiv
	Jung et al. [106]	2022	CVPR
	Diana et al. [107]	2022	FAccT
	Chakraborty et al. [92]	2022	FairWARE
	Wu et al. [108]	2022	CLeaR
	Suriyakumar et al. [109]	2022	arXiv

TABLE 4: Publications on Pre-processing bias mitigation methods - Part 2.

Category	Authors [Ref]	Year	Venue
Representation	Zemel et al. [110]	2013	ICML
	Edwards and Storkey [111]	2015	arXiv
	Louizos et al. [112]	2016	ICLR
	Xie et al. [113]	2017	NeurIPS
	Hacker and Wiedemann [114]	2017	arXiv
	McNamara et al. [115]	2017	arXiv
	Pérez-Suay et al. [116]	2017	ECML PKDD
	Calmon et al. [117]	2017	NeurIPS
	Komiyama and Shimao [118]	2017	arXiv
	Samadi et al. [119]	2018	NeurIPS
	Madras et al. [120]	2018	ICML
	du Pin Calmon et al. [121]	2018	IEEE J Sel
	Moyer et al. [122]	2018	NeurIPS
	Quadrianto et al. [123]	2018	arXiv
	Grgić-Hlača et al. [124]	2018	AAAI
	Song et al. [125]	2019	AISTATS
	Wang and Huang [126]	2019	arXiv
	Lahoti et al. [127]	2019	VLDB
	Feng et al. [128]	2019	arXiv
	Lahoti et al. [129]	2019	ICDE
	Creager et al. [130]	2019	ICML
	Gordaliza et al. [131]	2019	ICML
	Quadrianto et al. [132]	2019	CVPR
	Zhao et al. [133]	2020	ICLR
	Zehlike et al. [134]	2020	Data Min. Knowl. Discov
	Sarhan et al. [135]	2020	ECCV
	Tanu et al. [136]	2020	AISTATS
	Jaiswal et al. [137]	2020	AAAI
	Madhavan and Wadhwa [138]	2020	CIKM
	Ruoss et al. [139]	2020	NeurIPS
	Kim and Cho [140]	2020	AAAI
	Fong et al. [141]	2021	arXiv
	Salazar et al. [142]	2021	VLDB
	Gupta et al. [143]	2021	AAAI
	Grari et al. [144]	2021	ECML PKDD
	Zhu et al. [145]	2021	ICCV
	Oh et al. [146]	2022	arXiv
	Agarwal and Deshpande [147]	2022	FAccT
	Wu et al. [148]	2022	arXiv
	Shui et al. [149]	2022	arXiv
	Qi et al. [150]	2022	arXiv
	Balunović et al. [151]	2022	ICLR
	Kairouz et al. [152]	2022	IEEE Trans. Inf. Forensics Secur
	Liu et al. [153]	2022	Neural Process. Lett.
	Cerrato et al. [154]	2022	arXiv
	Kamani et al. [155]	2022	Mach. Learn.
	Rateike et al. [156]	2022	FAccT
	Galhotra et al. [157]	2022	SIGMOD
	Kim and Cho [158]	2022	Neurocomputing

then modified to reduce discrimination. The debiased causal network is used to generate a new dataset.

Sharma et al. [66] created additional data for augmentation by duplicating existing datasets and swapping the protected attribute of each instance. The newly-created data is successively added to the existing dataset.

4.1.3 Latent variables

Latent variable describes the augmentation of the training data with additional features that are preferably unbiased. In previous work, latent variables have been used to represent labels [101], [102] and group memberships (i.e., protected or unprotected group) [92], [98], [100], [103], [104], [105], [106], [107], [109].

For instance, Calders and Verwer [96] clustered the instances to detect those that should receive a positive latent label and those that should receive a negative one. For this purpose, they used an expectation maximization algorithm.

Gupta et al. [98] tackled the problem of bias mitigation for situations where group labels are missing in the datasets. To combat this issue, they created a latent "proxy" variable for the group membership and incorporated constraints for achieving fairness for such proxy groups in the training procedure.

Frequently, latent variables are considered when dealing with causal graphs [97], [99], [103].

4.1.4 Representation

Representation learning aims at learning a transformation of training data such that bias is reduced while maintaining as much information as possible.

The first bias mitigation approach for learning fair representations was Learning Fair Representations (LFR), proposed by Zemel et al. [110]. LFR translates representation learning into an optimization problem with two objectives: 1) removing information about the protected attribute; 2) minimizing the information loss of non-sensitive attributes.

A popular used approach for generating fair representations is optimization [114], [115], [117], [121], [122], [125], [127], [129], [131], [134], [149]. Other used techniques are:

- adversarial learning [111], [113], [120], [128], [133], [137], [139], [140], [144], [145], [150], [152];
- variational autoencoders [112], [130], [146], [153], [156];
- adversarial variational autoencoder [148];
- normalizing flows [151], [154];
- dimensionality reduction [116], [119], [136], [155];
- residuals [118];
- contrastive learning [143];
- neural style transfer [123], [132].

Another method for improving the fairness of the data representation is the removal [124], [126], [138] or addition of features [141], [142], [157]. Grgić-Hlača et al. [124] investigated fairness while using different sets of features, thereby making training features choices. Madhavan and Wadhwa [138] removed discriminating features from the training data. Salazar et al. [142] applied feature creation techniques, which apply nonlinear transformation, and then drop biased features.

4.2 In-processing Bias Mitigation Methods

This section presents in-processing methods; methods that mitigate bias during the training procedure of the algorithm. Overall, we found a total of 212 publications (see Table 5, Table 6 and Table 7 for more details) that apply in-processing methods. For more details on in-processing methods, we refer to the survey by Wan et al. [344], which provides information on 38 in-processing approaches developed for various ML tasks.

4.2.1 Regularization and Constraints

Regularization and constraints are both approaches that apply changes to the learning algorithm's loss function. Regularization adds a term to the loss function. While the original loss function is based on accuracy metrics, the purpose of regularization term is to penalize discrimination (i.e., discrimination leads to a higher loss of the ML algorithm. Constraints on the other hand determine specific bias levels (according to loss functions) that cannot be breached during training.

To widen the range of fairness definitions that can be considered when applying constraints, Celis et al. [261] proposed a Meta-algorithm. This Meta-algorithm takes a fairness constraint as input.

When applied to Decision Trees, regularization can be used to modify the splitting criteria [160], [173], [176], [188], [189], [198], [201]. Traditionally, leaves are iteratively split to achieve an improvement in accuracy. To improve fairness while training, Kamiran et al. [160] considered fairness in addition to accuracy when leaf splitting. They applied three splitting strategies:

1) only allow non-discriminatory splits;
2) choose best split according to $\delta_{accuracy}/\delta_{discrimination}$;
3) 3) choose best split according to $\delta_{accuracy} + \delta_{discrimination}$.

TABLE 5: Publications on In-processing bias mitigation methods.

Category	Authors [Ref]	Year	Venue
Regularization	Kamiran et al. [160]	2010	ICDM
	Kamishima et al. [161]	2011	ICDMW
	Kamishima et al. [162]	2012	ECML PKDD
	Ristanoski et al. [163]	2013	CIKM
	Fish et al. [164]	2015	FATML
	Berk et al. [165]	2017	arXiv
	Pérez-Suay et al. [116]	2017	ECML PKDD
	Bechavod and Ligett [166]	2017	arXiv
	Quadrianto and Sharmanska [167]	2017	NeurIPS
	Raff et al. [168]	2018	AIES
	Goel et al. [169]	2018	AAAI
	Enni and Assent [170]	2018	ICDM
	Mary et al. [171]	2019	ICML
	Beutel et al. [172]	2019	AIES
	Zhang et al. [173]	2019	ICDMW
	Aghaei et a l. [174]	2019	AAAI
	Huang and Vishnoi [175]	2019	ICML
	Zhang and Ntoutsi [176]	2019	IJCAI
	Tavakol [177]	2020	SIGIR
	Baharlouei et al. [178]	2020	ICLR
	Di Stefano et al. [179]	2020	arXiv
	Kim et al. [180]	2020	ICML
	Jiang et al. [181]	2020	UAI
	Romano et al. [182]	2020	NeurIPS
	Ravichandran et al. [183]	2020	arXiv
	Liu et al. [184]	2020	Preprint
	Keya et al. [185]	2020	arXiv
	Hickey et al. [186]	2020	ECML PKDD
	Kamani [187]	2020	Thesis
	Abay et al. [73]	2020	arXiv
	Chuang and Mroueh [76]	2021	ICLR
	Zhang and Weiss [188]	2021	ICDM
	Ranzato et al. [189]	2021	CIKM
	Kang et al. [190]	2021	arXiv
	Grari et al. [191]	2021	IJCAI
	Wang et al. [192]	2021	SIGKDD
	Mishler and Kennedy [193]	2021	arXiv
	Lowy et al. [194]	2021	arXiv
	Zhao et al. [195]	2021	arXiv
	Yurochkin and Sun [196]	2021	ICLR
	Sun et al. [44]	2022	EuroS&P
	Zhao et al. [197]	2022	WSDM
	Wang et al. [198]	2022	CAV
	Deng et al. [199]	2022	arXiv
	Lee et al. [200]	2022	Entropy
	Zhang and Weiss [201]	2022	AAAI
	Jiang et al. [202]	2022	ICLR
	Lee et al. [203]	2022	ICASSP
	Do et al. [204]	2022	ICML
	Patil and Purcell [205]	2022	Future Internet
	Kim and Cho [158]	2022	Neurocomputing
Adversarial	Beutel et al. [206]	2017	arXiv
	Gillen et al. [207]	2018	NeurIPS
	Kearns et al. [208]	2018	ICML
	Wadsworth et al. [209]	2018	arXiv
	Agarwal et al. [210]	2018	ICML
	Raff and Sylvester [211]	2018	DSAA
	Zhang et al. [212]	2018	AIES
	Sadeghi et al. [213]	2019	ICCV
	Adel et al. [214]	2019	AAAI
	Zhao and Gordon [215]	2019	NeurIPS
	Celis and Keswani [216]	2019	nan
	Beutel et al. [172]	2019	AIES
	Grari et al. [217]	2019	ICDM
	Xu et al. [63]	2019	Big Data
	Yurochkin et al. [218]	2020	ICLR
	Garcia de Alford et al. [219]	2020	SMU DSR
	Roh et al. [220]	2020	ICML
	Delobelle et al. [221]	2020	ASE
	Rezaei et al. [222]	2020	AAAI
	Lahoti et al. [223]	2020	NeurIPS
	Amend and Spurlock [77]	2021	JCSC
	Rezaei et al. [224]	2021	AAAI
	Grari et al. [191]	2021	IJCAI
	Grari et al. [103]	2021	arXiv
	Liang et al. [105]	2022	arXiv
	Chen et al. [104]	2022	arXiv
	Tao et al. [225]	2022	FSE
	Petrović et al. [226]	2022	Neurocomputing
	Yang et al. [227]	2022	medRxiv
	Yazdani-Jahromi et al. [228]	2022	arXiv

TABLE 6: Publications on In-processing bias mitigation methods - Part 2.

Category	Authors [Ref]	Year	Venue
Constraints	Dwork et al. [229]	2012	ITCS
	Calders et al. [230]	2013	ICDM
	Fukuchi and Sakuma [231]	2015	arXiv
	Fukuchi et al. [232]	2015	IEICE Trans. Inf.& Syst.
	Goh et al. [233]	2016	NeurIPS
	Zafar et al. [234]	2017	AISTATS
	Russel et al. [235]	2017	NeurIPS
	Corbett-Davies et al. [236]	2017	KDD
	Quadrianto and Sharmanska [167]	2017	NeurIPS
	Zafar et al. [237]	2017	TheWebConf
	Komiyama and Shimao [118]	2017	arXiv
	Woodworth et al. [238]	2017	COLT
	Kilbertus et al. [97]	2017	NeurIPS
	Zafar et al. [239]	2017	NeurIPS
	Gillen et al. [207]	2018	NeurIPS
	Olfat and Aswani [240]	2018	AISTATS
	Narasimhan [241]	2018	AISTATS
	Kearns et al. [208]	2018	ICML
	Zhang and Bareinboim [242]	2018	AAAI
	Heidari et al. [243]	2018	NeurIPS
	Kim et al. [244]	2018	NeurIPS
	Gupta et al. [98]	2018	arXiv
	Agarwal et al. [210]	2018	ICML
	Farnadi et al. [245]	2018	AIES
	Goel et al. [169]	2018	AAAI
	Nabi and Shpitser [246]	2018	AAAI
	Wu et al. [247]	2018	arXiv
	Zhang and Bareinboim [248]	2018	NeurIPS
	Grgić-Hlača et al. [124]	2018	AAAI
	Komiyama et al. [249]	2018	ICML
	Donini et al. [250]	2018	NeurIPS
	Balashankar et al. [251]	2019	arXiv
	Zafar et al. [252]	2019	JMLR
	Lamy et al. [253]	2019	NeurIPS
	Cotter et al. [254]	2019	ALT
	Jung et al. [255]	2019	arXiv
	Oneto et al. [100]	2019	AIES
	Cotter et al. [256]	2019	J. Mach. Learn. Res.
	Wick et al. [257]	2019	NeurIPS
	Cotter et al. [258]	2019	ICML
	Nabi et al. [259]	2019	ICML
	Xu et al. [260]	2019	TheWebConf
	Celis et al. [261]	2019	FAccT
	Agarwal et al. [262]	2019	ICML
	Kilbertus et al. [263]	2020	AISTATS
	Lohaus et al. [264]	2020	ICML
	Ding et al. [265]	2020	AAAI
	Chzhen et al. [266]	2020	NeurIPS
	Wang et al. [267]	2020	NeurIPS
	Cho et al. [268]	2020	NeurIPS
	Oneto et al. [269]	2020	IJCNN
	Maity et al. [270]	2020	arXiv
	Chzhen and Schreuder [271]	2020	arxiv
	Manisha and Gujar [272]	2020	IJCAI
	Scutari et al. [273]	2021	arXiv
	Celis et al. [274]	2021	NeurIPS
	Celis et al. [275]	2021	PMLR
	Petrović et al. [276]	2021	Eng. Appl. Artif. Intell.
	Padh et al. [277]	2021	Uncertainty artif. intell.
	Zhao et al. [278]	2021	KDD
	Zhang et al. [279]	2021	MOD
	Li et al. [280]	2021	LAK
	Du and Wu [82]	2021	CIKM
	Perrone et al. [281]	2021	AIES
	Słowik and Bottou [282]	2021	arXiv
	Mishler and Kennedy [193]	2021	arXiv
	Lawless et al. [283]	2021	arXiv
	Choi et al. [284]	2021	AAAI
	Park et al. [285]	2022	WWW
	Wang et al. [198]	2022	CAV
	Zhao et al. [286]	2022	KDD
	Boulitsakis-Logothetis [287]	2022	arXiv
	Hu et al. [288]	2022	arXiv
	Wu et al. [108]	2022	CLeaR
Compositional	Calders and Verwer [96]	2010	Data Min. Knowl. Discov
	Pleiss et al. [289]	2017	NeurIPS
	Dwork et al. [290]	2018	FAccT
	Ustun et al. [291]	2019	ICML
	Oneto et al. [100]	2019	AIES
	Iosifidis et al. [64]	2019	Big Data
	Monteiro and Reynoso-Meza [292]	2021	PLM
	Ranzato et al. [189]	2021	CIKM
	Mishler and Kennedy [193]	2021	arXiv
	Kobayashi and Nakao [293]	2021	DiTTEt
	Jin et al. [294]	2022	ICML
	Chen et al. [90]	2022	FSE
	Roy et al. [295]	2022	DS
	Liu and Vicente [296]	2022	CMS
	Blanzeisky and Cunningham [297]	2022	Knowl Eng Rev
	Boulitsakis-Logothetis [287]	2022	arXiv
	Suriyakumar et al. [109]	2022	arXiv

While constraints and regularization usually utilize group fairness definitions, they have also been applied for achieving individual fairness [207], [229], [244], [255]. Moreover, they can be applied to achieve fairness for multiple sensitive attributes and fairness definitions [177], [190], [190], [208], [249], [277], or extend existing adjustments, such as adding fairness regularization in addition to the L2 norm, which is used to avoid overfitting [161], [162].

TABLE 7: Publications on In-processing bias mitigation methods - Part 3.

Category	Authors [Ref]	Year	Venue
Adjusted	Luo et al. [298]	2015	DaWaK
	Joseph et al. [299]	2016	NeurIPS
	Johnson et al. [300]	2016	Stat Sci
	Kusner et al. [301]	2017	NeurIPS
	Joseph et al. [302]	2018	AIES
	Hashimoto et al. [303]	2018	ICML
	Hébert-Johnson et al. [304]	2018	ICML
	Chiappa and Isaac [305]	2018	IFIP
	Alabi et al. [306]	2018	COLT
	Madras et al. [307]	2018	NeurIPS
	Kamishima et al. [308]	2018	Data Min Knowl Discov
	Kilbertus et al. [309]	2018	ICML
	Dimitrakakis et al. [310]	2019	AAAI
	Chakraborty et al. [311]	2019	arXiv
	Noriega-Campero et al. [312]	2019	AIES
	Chiappa [313]	2019	AAAI
	Madras et al. [99]	2019	FAccT
	Iosifidis and Ntoutsi [314]	2019	CIKM
	Kilbertus et al. [263]	2020	AISTATS
	Zhang and Ramesh [315]	2020	arXiv
	Chakraborty et al. [67]	2020	FSE
	Mandal et al. [316]	2020	NeurIPS
	Hu et al. [69]	2020	DS
	Liu et al. [184]	2020	Preprint
	da Cruz [317]	2020	Thesis
	Iosifidis and Ntoutsi [318]	2020	DS
	Kamani [187]	2020	Thesis
	Martinez et al. [319]	2020	ICML
	Ignatiev et al. [320]	2020	CP
	Ezzeldin et al. [321]	2021	arXiv
	Zhang et al. [75]	2021	PAKDD
	Wang et al. [322]	2021	FAccT
	Ozdayi et al. [323]	2021	arXiv
	Islam et al. [324]	2021	AIES
	Sharma et al. [325]	2021	AIES
	Cruz et al. [79]	2021	ICDM
	Lee et al. [326]	2021	ICML
	Hort and Sarro [327]	2021	ASE
	Perrone et al. [281]	2021	AIES
	Roh et al. [328]	2021	ICLR
	Valdivia et al. [329]	2021	Int. J. Intell. Syst.
	Wang et al. [330]	2022	arXiv
	Roy and Ntoutsi [331]	2022	ECML PKDD
	Sikdar et al. [332]	2022	FAccT
	Agarwal and Deshpande [147]	2022	FAccT
	Park et al. [285]	2022	WWW
	Djebrouni [333]	2022	Eurosys
	Short and Mohler [334]	2022	Int. J. Forecast.
	Maheshwari and Perrot [335]	2022	arXiv
	Zhao et al. [286]	2022	KDD
	Tizpaz-Niari et al. [336]	2022	ICSE
	Roy et al. [295]	2022	DS
	Mohammadi et al. [337]	2022	arXiv
	Gao et al. [338]	2022	ICSE
	Huang et al. [339]	2022	Expert Syst. Appl.
	Candelieri et al. [340]	2022	arXiv
	Anahideh et al. [341]	2022	Expert Syst. Appl.
	Rateike et al. [156]	2022	FAccT
	Li et al. [342]	2022	arXiv
	Iosifidis et al. [343]	2022	KAIS

4.2.2 Adversarial Learning

Adversarial learning simultaneously trains classification models and their adversaries [345]. While the classification model is trained to predict ground truth values, the adversary is trained to exploit fairness issues. Both models then perform against each other, to improve their performance.

Zhang et al. [212] trained a Logistic Regression model to predict the label Y while preventing an adversary from predicting the protected attribute under consideration of three fairness metrics: Demographic Parity, Equality of Odds, and Equality of Opportunity. Both, predictor and adversary, are implemented as Logistic regression models.

Similarly, Beutel et al. [206] trained a neural network to predict two outputs: labels and sensitive attributes. While a high overall accuracy is desired, the adversarial setting optimises a low ability to predict sensitive information. The network is designed to share layers between the two output, such that only one model is trained [172], [211], [213], [214], [221].

Lahoti et al. [223] proposed Adversarially Reweighted Learning (ARL) in which a learner is trained to optimize performance on a classification task while the adversary adjusts the weights of computationally-identifiable regions in the input space with high training loss. By so-doing, the learner can then improve performance in these regions.

Other than using adversaries to prevent the ability to predict sensitive attributes (e.g., for reducing bias according to population groups), it has also been used to improve robustness to data poisoning [220], to improve individual fairness [218], and to reweigh training data [226]. In particular, Petrović et al. [226] used adversarial training to learn a reweighing function for training data instances as an in-processing procedure (contrary to applying reweighing as pre-processing, see Section 4.1.2).

4.2.3 Compositional

Compositional approaches combat bias by training multiple classification models. Predictions can then be made by a specific classification model for each population group (e.g., privileged and unprivileged) [96], [100], [109], [287], [289], [291], [294] or in an ensemble fashion (i.e., a voting of multiple classification models at the same time) [64], [90], [189], [193], [293], [295], [296], [346].

While decoupled classification models for privileged and unprivileged groups can achieve improved accuracy for each group, the amount training data for each classifier is reduced. To reduce the impact of small training data sizes Dwork et al. [290] utilised transfer training. With their transfer learning approach, they trained classifiers on data for the respective group and data from the other groups with reduced weight. Ustun et al. [291] built upon the work of Dwork et al. [290] and incorporates "preference guarantees", which states that each group prefers their decoupled classifier over a classifier trained on all training data and any classifier of the other groups. Similarly, Suriyakumar et al. [109] followed the concept of "fair use", which states that if a classification uses sensitive group information, it should improve performance for every group.

Training multiple classification models with different fairness goals allows for the creation of a pareto-front of solutions [193], [295], [296], [297], [329]. Practitioners can then choose which fairness-accuracy trade-off best suits their need. For example, Liu and Vicente [296] treated bias mitigation as multi-objective optimization problem that explores fairness-accuracy trade-offs under consideration of multiple fairness metrics. Mishler and Kennedy [193] proposed an ensemble method that builds classification models based on a weighted combination of metrics chosen by users.

4.2.4 Adjusted Learning

Adjusted learning methods mitigate the bias via changing the learning procedure of algorithms or the creation of novel algorithms [18].

Changes have been suggested for a variety of classification models, including Bayesian models [310], [347], Markov Random Fields [315], Neural Networks [69], [211], [319], Decision Trees, bandits [299], [302], [348], boosting [295], [304], [314], [318], Logistic Regression [328]. We outline a selection of publications in the following, to provide insight on techniques applied to different classification models.

Noriega-Campero et al. [312] proposed an active learning framework for training Decision Trees. During the training, a decision maker is able to collect more information about individuals to achieve fairness in predictions. In this context, not all information about individuals is available. There is an information budget that determines how many enquiries can be performed. Similarly, Anahideh et al. [341] used an active learning framework to balance accuracy and fairness by selecting instances to be labelled.

Madras et al. [307] proposed a rejection learning approach for joint decision-making with classification models and external decision makers. In particular, the classification model learns when to defer from making prediction (i.e., when it is more useful to have predictions from external decision makers). If the coverage of classification can be reduced (i.e., the classification model abstains from making some of the predictions), selective classification approaches can be used [326].

Martinez et al. [319] proposed the algorithm Approximate Projection onto Star Sets (APStar) to train Deep Neural Networks to minimize the maximum risk among all population groups. This procedure ensures that the final classifier is part of the Pareto Front [349]. Hu et al. [69] incorporated representation learning into the training procedure of Neural Networks to learn them jointly the classifier.

Hébert-Johnson et al. [304] proposed *Multicalibration*, a learning procedure similar to boosting. A classifier is trained iteratively. At each iteration, the predictions of the most biased subgroup are corrected until the classifier is adequately calibrated.

Hashimoto et al. [303] found fairness issues with the use of empirical risk minimization and proposed the use of distributionally robust optimization (DRO) when training classifiers such as Logistic Regression. During training, DRO optimizes the worst-case risk over all groups present.

Kilbertus et al. [309] adjusted the training procedure for Logistic Regression to take privacy into account. Sensitive user information is encrypted such that it cannot be used for classification tasks while retaining the ability to verify fairness issues. By doing so, users can provide sensitive information without the fear that someone can read them.

The learning procedure of existing classification models has also been adjusted by tuning their hyper-parameters [67], [79], [281], [311], [317], [324], [327], [329], [336].

4.3 Post-processing Bias Mitigation Methods

Post-processing bias mitigation methods are applied once a classification model has been successfully trained. With 56 publications that apply post-processing methods (Table 8), post-processing methods are the least frequently applied of those covered in this survey.

TABLE 8: Publications on Post-processing bias mitigation methods.

Category	Authors [Ref]	Year	Venue
Input	Adler et al. [350]	2018	KAIS
	Li et al. [53]	2022	ICSE
Classifier	Calders and Verwer [96]	2010	Data Min. Knowl. Discov
	Kamiran et al. [160]	2010	ICDM
	Hardt et al. [351]	2016	NeurIPS
	Woodworth et al. [238]	2017	COLT
	Pleiss et al. [289]	2017	NeurIPS
	Gupta et al. [98]	2018	arXiv
	Morina et al. [352]	2019	arXiv
	Noriega-Campero et al. [312]	2019	AIES
	Kim et al. [353]	2019	AIES
	Kanamori and Arimura [354]	2019	JSAI
	Kim et al. [180]	2020	ICML
	Jiang et al. [181]	2020	UAI
	Savani et al. [355]	2020	NeurIPS
	Chzhen et al. [356]	2020	NeurIPS
	Chzhen et al. [266]	2020	NeurIPS
	Awasthi et al. [357]	2020	PMLR
	Chzhen and Schreuder [271]	2020	arxiv
	Schreuder and Chzhen [358]	2021	UAI
	Kanamori and Arimura [359]	2021	JSAI
	Mishler et al. [360]	2021	FAccT
	Mishler and Kennedy [193]	2021	arXiv
	Du et al. [361]	2021	NeurIPS
	Grabowicz et al. [362]	2022	FAccT
	Zhang et al. [363]	2022	FairWARE
	Mehrabi et al. [364]	2022	TrustNLP
	Wu and He [365]	2022	FAccT
	Marcinkevics et al. [366]	2022	MLHC
	Iosifidis et al. [343]	2022	KAIS
Output	Pedreschi et al. [367]	2009	SDM
	Kamiran et al. [9]	2012	ICDM
	Fish et al. [164]	2015	FATML
	Fish et al. [368]	2016	SDM
	Kim et al. [244]	2018	NeurIPS
	Zhang et al. [42]	2018	IJCAI
	Menon and Williamson [369]	2018	FAccT
	Liu et al. [370]	2018	arXiv
	Kamiran et al. [10]	2018	J. Inf. Sci.
	Chiappa [313]	2019	AAAI
	Chzhen et al. [371]	2019	NeurIPS
	Iosifidis et al. [64]	2019	Big Data
	Lohia et al. [372]	2019	ICASSP
	Wei et al. [101]	2020	PMLR
	Alabdulmohsin [373]	2020	arXiv
	Alabdulmohsin and Lucic [374]	2021	NeurIPS
	Nguyen et al. [375]	2021	J. Inf. Sci.
	Kobayashi and Nakao [293]	2021	DiTTEt
	Lohia [376]	2021	arXiv
	Jang et al. [377]	2022	AAAI
	Pentyala et al. [87]	2022	arXiv
	Snel and van Otterloo [378]	2022	Com. Soc. Res. J.
	Alghamdi et al. [379]	2022	arXiv
	Mohammadi et al. [337]	2022	arXiv
	Zeng et al. [380]	2022	arXiv
	Zeng et al. [381]	2022	arXiv

4.3.1 Input Correction

Input correction approaches apply a modification step to the testing data. This is comparable to pre-processing approaches (Section 4.1) [18], which conduct modifications to training data (e.g., relabelling, perturbation and representation learning).

We found only two publications that apply input corrections to testing data, both of which use perturbations. While Adler et al. [350] used perturbation in a post-processing stage, Li et al. [53] first performed perturbation in a pre-processing stage and then applied an identical procedure for post-processing.

4.3.2 Classifier Correction

Post-processing approaches can also directly be applied to classification models, which Savani et al. [355] called intra-processing. A successfully trained classification model is adapted to obtain a fairer one. Such modification have been applied to Naive Bayes [96], Logistic Regression [181], Decision Trees [160], [359], [363], Neural Networks [355], [361], [364], [366] and Regression Models [266].

Hardt et al. [351] proposed the modification of classifiers to achieve fairness with respect to Equalized Odds and Equality of Opportunity. Given an unfair classifier $\widehat{Y}$, the classifier $\widetilde{Y}$ is derived by solving an optimization problem under consideration of fairness loss terms. This approach has been adapted and extended by further publications [98], [352], [357], [360].

Woodworth et al. [238] showed that this kind of modification can lead to a poor accuracy, for example when the loss function is not strictly convex. In addition to constraints during training, they proposed an adaptation of the approach by Hardt et al. [351].

Pleiss et al. [289] split a classifier in two (h_0, h_1, for the privileged and unprivileged group). To balance the false positive and false negative rate of the two classifiers, h_1 is adjusted such that with a probability of α the class mean is returned rather than the actual predication.Noriega-Campero et al. [312] followed the calibration approach of Pleiss et al. [289].

Kamiran et al. [160] modified Decision Tree classifiers by relabeling leaf nodes. The goal of relabeling was to reduce bias while sacrificing as little accuracy as possible. A greedy procedure was followed which iteratively selects the best leaf to relabel (i.e., highest ratio of fairness improvement per accuracy loss). Kanamori and Arimura [359] formulated the modification of branching thresholds for Decision Trees as a mixed integer program.

Kim et al. [353] proposed *Multiaccuracy Boost*, a post-processing approach similar to boosting for training classifiers. Given a black-box classifier and a learning algorithm, *Multiaccuracy Boost* iteratively adapts the current classifier based on its predictive performance.

4.3.3 Output Correction

The latest stage of applying bias mitigation methods is the correction of the output. In particular, the predicted labels are modified.

Pedreschi et al. [367] considered the correction of rule based classifiers, such as CPAR [382]. For each individual, the k rules with highest confidence are selected to determine the probability for each output label. Given that some of the rules can be discriminatory, their confidence level is adjusted to reduce biased labels.

Menon and Williamson [369] proposed a plugin approach for thresholding predictions. To determine the thresholds to use, the class probabilities are estimated using logistic regression.

Kamiran et al. [9], [10] introduced the notion of reject option which modifies the prediction of individuals close to the decision boundary. In particular, individuals belonging to the unprivileged group receive a positive outcome and privileged individuals an unfavourable outcome. Similarly, Lohia et al. [372] relabeled individuals that are likely to receive biased outcomes, but rather than considering the decision boundary, they used an "individual bias detector" to find predictions that are likely suffer from individual discrimination. This work was extended in 2021, where individuals were ranked based on their "Unfairness Quotient" (i.e., the difference between regular prediction and with perturbed protected attribute). Fish et al. [368] proposed a confidence-based approach which returns a positive label for each prediction above a given threshold. This has also been applied to AdaBoost [164]. Other than using a general threshold for all instances, group dependent thresholds can be used [64], [87], [293], [371], [373], [377], [380], [381].

Chiappa [313] addressed the fairness of causal models under consideration of a counterfactual world in which individuals belong to a different population group. The impact of the protected attribute on the prediction outcome is corrected to ensure that it coincides with counterfactual predictions. This way, sensitive information is removed while other information remains unchanged.

4.4 Combined Approaches

While most publications propose the use of a single type of bias mitigation method, we found 70 that applied multiple techniques at the same time (e.g., two pre-processing methods, one in-processing and one post-processing methods). Table 9 summarizes these approaches.

Among these 70 publications, 86% (60 out of 70) applied in-processing, 54% (38 out of 70) applied pre-processing, and 31% (22 out of 70) applied post-processing methods.

Additionally, 26 out of 70 publications applied multiple types of bias mitigation methods but at the same stage of the development process (e.g., two pre-processing approaches). In particular, the are 7 publications which applied multiple pre-processing methods. Among these 7 publications, 5 applied sampling and relabeling [37], [39], [41], [43], [44]. The remaining 19 out of 26 publications applied multiple in-processing methods, 17 of which include regularization or constraints.

47 publications applied at least two methods at different stages of the development process for ML models (e.g. one pre-processing and one in-processing method). This illustrates that bias mitigation methods can be used in conjunction [383]. Moreover, there are three publications that addressed bias mitigation at each stage: pre-processing, in-processing and post-processing [64], [96], [98].

Calders and Verwer [96] proposed three approaches for achieving discrimination-free classification of naive bayes models. At first, a latent variable is added to represent unbiased labels. The data is then used to train a model for each possible sensitive attribute value. Lastly, the probabilities output by the model are modified to account for unfavourable treatment (i.e., increasing the probability of positive outcomes for the unprivileged group and reducing it for the privileged group).

Gupta et al. [98] tackled the problem of bias mitigation for situation where group labels are missing in the datasets. To combat this issue, they created a latent "proxy" variable for the group membership and incorporated constraints for achieving fairness for such proxy groups in the training procedure. Lastly, they followed the approach of Hardt et al. [351] to debias and existing classifier by adding an additional variable to the prediction problem (see Section 4.3.2).

Iosifidis et al. [64] followed an ensemble approach of multiple AdaBoost classifiers. In particular, each classifier is trained on an equal amount of instances from each population group and label by sampling. Predictions are then modified by applying group-dependent thresholds.

TABLE 9: Publications with multiple bias mitigation methods. "X" indicates that the publication applies a bias mitigation approach of the corresponding category (i.e., pre-, in-, or post-processing).

Authors [Ref]	Processing Method		
	Pre	In	Post
Sun et al. [44]	x x	x	
Calders et al. [37]	x x		
Žliobaite et al. [39]	x x		
Hajian et al. [40]	x x		
Kamiran and Calders [41]	x x		
Iosifidis et al. [43]	x x		
Chakraborty et al. [92]	x x		
Oneto et al. [100]	x	x x	
Calders and Verwer [96]	x	x	x
Gupta et al. [98]	x	x	x
Iosifidis et al. [64]	x	x	x
Pérez-Suay et al. [116]	x	x	
Komiyama and Shimao [118]	x	x	
Kilbertus et al. [97]	x	x	
Grgić-Hlača et al. [124]	x	x	
Madras et al. [99]	x	x	
Xu et al. [63]	x	x	
Abay et al. [73]	x	x	
Hu et al. [69]	x	x	
Chakraborty et al. [67]	x	x	
Chuang and Mroueh [76]	x	x	
Zhang et al. [75]	x	x	
Grari et al. [103]	x	x	
Du and Wu [82]	x	x	
Amend and Spurlock [77]	x	x	
Cruz et al. [79]	x	x	
Chen et al. [104]	x	x	
Liang et al. [105]	x	x	
Agarwal and Deshpande [147]	x	x	
Chen et al. [90]	x	x	
Wu et al. [108]	x	x	
Rateike et al. [156]	x	x	
Kim and Cho [158]	x	x	
Suriyakumar et al. [109]	x	x	
Zhang et al. [42]	x		x
Wei et al. [101]	x		x
Pentyala et al. [87]	x		x
Li et al. [53]	x		x
Mishler and Kennedy [193]		x x x	x
Quadrianto and Sharmanska [167]		x x	
Agarwal et al. [210]		x x	
Gillen et al. [207]		x x	
Kearns et al. [208]		x x	
Goel et al. [169]		x x	
Beutel et al. [172]		x x	
Kilbertus et al. [263]		x x	
Liu et al. [184]		x x	
Kamani [187]		x x	
Perrone et al. [281]		x x	
Grari et al. [191]		x x	
Ranzato et al. [189]		x x	
Park et al. [285]		x x	
Wang et al. [198]		x x	
Zhao et al. [286]		x x	
Roy et al. [295]		x x	
Boulitsakis-Logothetis [287]		x x	
Kamiran et al. [160]		x	x
Fish et al. [164]		x	x
Woodworth et al. [238]		x	x
Pleiss et al. [289]		x	x
Kim et al. [244]		x	x
Chiappa [313]		x	x
Noriega-Campero et al. [312]		x	x
Chzhen and Schreuder [271]		x	x
Kim et al. [180]		x	x
Jiang et al. [181]		x	x
Chzhen et al. [266]		x	x
Kobayashi and Nakao [293]		x	x
Iosifidis et al. [343]		x	x
Mohammadi et al. [337]		x	x

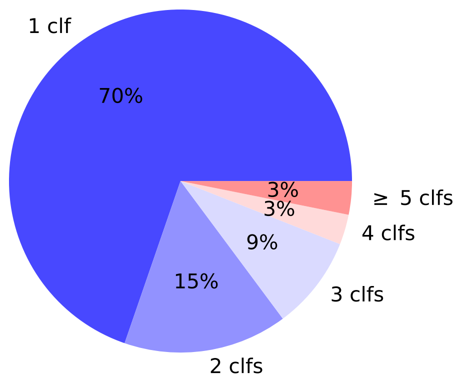


Fig. 2: Number of classification models (clf) used for evaluation.

4.5 Classification Models

Here we outline the classification models on which the three types of bias mitigation methods (pre-, in-, post-processing) have been applied on. Table 10 shows the frequency with which each type of classification model has been applied.

Currently, the most frequently used classification model is Logistic Regression, for each method type (pre-, in-, post-processing), with a total of 140 unique publications using it for their experiments. The next most frequently used classification models are Neural Networks (NN). A total of 102 publication used NNs for their experiments, with the majority being in-processing methods. Linear Regression models have been used in 22 publications.

Decision Trees (36 publications) and Random Forests (45 publications) are also frequently used. Moreover, different Decision Tree variants have been used, such as Hoeffding trees, C4.5, J48 and Bayesian random forests.

While the range of classification models is diverse, some of them are similar to one another:

- Boosting: AdaBoost, XGBoost, SMOTEBoost, Boosting, LightGBM, OSBoost, Gradient Tree Boosting, CatBoost;
- Rule-based: RIPPER, PART, CBA, Decision Set, Rule Sets, Decision Rules.

Figure 2 illustrates the number of different classification models considered during experiments. It is clear to see that the majority of publications (70%) applied their bias mitigation method to only one classification model. While in-processing methods are model specific and directly modify the training procedure, pre-processing and most post-processing bias mitigation methods can be developed independently from the classification models they are used for. Therefore, they can be devised once and applied to multiple classification models for evaluating their performance. Our observations confirm this intuition: only 24% of publications with in-processing methods consider more than one classification model, while 35% and 43% of pre- and post-processing methods consider more than one respectively.

TABLE 10: Frequency of classification model usage for evaluating bias mitigation methods. Amounts are provided for each category and as a unique measure to avoid counting publications with multiple approaches double.

Model	Unique	Processing Method Pre	In	Post
Logistic Regression	140	58	80	19
NN	102	34	65	17
Random Forest	45	20	22	14
SVM	37	15	18	9
Decision Tree	36	14	16	9
Naive Bayes	24	12	11	5
Linear Regression	22	4	20	3
AdaBoost	8	1	5	4
XGBoost	8	1	6	1
Nearest Neighbour	7	3	2	3
Causal	7	2	6	1
Nearest Neighbor	6	4	0	2
LightGBM	4	2	3	0
Bandit	3	0	3	0
Boosting	3	0	2	2
J48	2	1	1	0
Bayesian	2	0	1	1
Hoeffding Tree	2	1	1	0
Gaussian Process	2	2	0	0
CPAR	1	0	0	1
RIPPER	1	1	0	0
PART	1	1	0	0
C4.5	1	1	0	0
CBA	1	0	1	0
Lattice	1	1	1	1
Lasso	1	0	1	0
PSL	1	0	1	0
BART	1	0	1	0
RTL	1	0	1	0
Tree Ensemble	1	0	1	0
AUE	1	1	0	0
CART	1	0	1	0
SMOTEBoost	1	0	1	0
Gradient boosted trees	1	1	0	1
Cox model	1	0	1	0
Decision Rules	1	0	1	0
Gradient Tree Boosting	1	0	1	0
Kmeans	1	0	1	0
OSBoost	1	0	1	0
POEM	1	0	1	0
Markov random filed	1	0	1	0
SMSGDA	1	0	1	0
Probabilistic circuits	1	0	1	0
Rule Sets	1	0	1	0
Ridge Regression	1	0	1	1
Extreme Random Forest	1	1	0	0
Factorization Machine	1	1	0	0
Discriminant analysis	1	0	1	0
Generalized Linear Model	1	0	1	0

5 Datasets

In this section, we investigate the use of datasets for evaluating bias mitigation methods. Among these datasets, some have been divided into multiple subsets (e.g., risk of recidivism or violent recidivism, medical data for different time periods). For clarity, we treat data from the same source as a single dataset.

Following this procedure, we gathered a total of 81 unique datasets. We discuss these datasets in Section 5.1 (e.g., what is the most frequently used dataset?) and Section 5.2 (e.g., how many datasets do experiments consider?). Additionally, 56 publications created synthetic or semi-synthetic datasets for their experiments. Section 5.3 provides information on

TABLE 11: Frequency of widely used datasets (i.e., used in at least three publications).

Dataset Name	Frequency	Percentage
Adult [385]	249	77%
COMPAS [2]	166	51%
German [385]	97	30%
Communities and Crime [386]	42	13%
Bank [387]	38	12%
Law School [388]	33	10%
Default [389]	24	7%
Dutch Census [390]	16	5%
Health [391]	14	4%
MEPS [392]	14	4%
Drug [393]	9	3%
Student [394]	8	2%
Heart disease [385]	7	2%
National Longitudinal Survey of Youth [395]	6	2%
SQF [396]	5	2%
Arrhythmia [385]	5	2%
Wine [397]	4	1%
Ricci [398]	4	1%
University Anonymous (UNIV)	3	1%
Home credit [399]	3	1%
ACS [384]	3	1%
MIMICIII [400]	3	1%

the creation of such synthetic data.

For further details on datasets, we refer to Le Quy et al. [26] who surveyed 15 datasets and provided detailed information on the features and dataset characteristics. Additionally, Kuhlman et al. [16] gathered 22 datasets from publications published in the ACM Fairness, Accountability, and Transparency (FAT) Conference and 2019 AAAI/ACM conference on Articial Intelligence, Ethics and Society (AIES).

5.1 Dataset Usage

In this section, we investigate the frequency with which each dataset set has been used. The purpose of this analysis is to highlight the importance of each dataset and recommend the most important datasets to use for evaluating bias mitigation methods.

Among the 81 datasets, two are concerned with synthetic data (i.e., "synthetic" and "semi-synthetic") which we address in Section 5.3. Therefore, we are left with 81 datasets. 59% of the datasets (48 out of 81) are used by only one publication during their experiments. Another 14% of the datasets (11 out of 81) are only used twice. Thereby, 73% of the datasets (59 out of 81) are used rarely (by one or two publications).

Table 11 list the frequency of the remaining 22 datasets (used in three or more publications). In addition to the frequency, a percentage is provided (i.e., how many of the 324 publications use this datasets). Among all datasets, the Adult dataset is used most frequently (by 77% of the publications). While the Adult dataset contains information from the 1994 US census, Ding et al. [384] derived new datasets from the US census from 2014 to 2018.

Five other datasets are used by 10% or more of the publications (COMPAS, German Communities and Crime, Bank, Law School). This shows that in order to enable a simple comparison with existing work, one should consider at least the Adult and COMPAS dataset. A list of all datasets can be found in our online repository [22].

5.2 Dataset Count

In addition to detecting the most popular datasets for evaluating bias mitigation methods, we investigate the number of different datasets used, as this impacts the diversity of the performance evaluation [16].

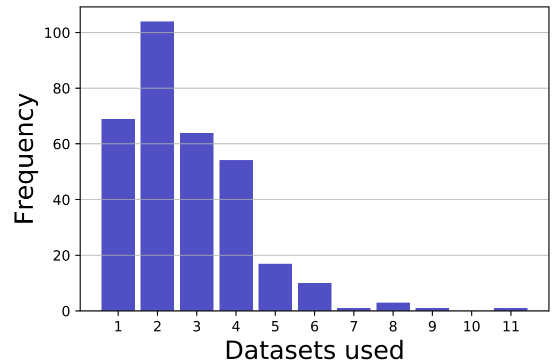


Fig. 3: Number of datasets used per publication.

Figure 3 visualizes the number of datasets used for each of the 324 publications.

The most commonly used number of datasets considered for experiments is two, which has been observed in 104 out 324 of the publications. Over all, it can be seen that the number of considered datasets is relatively small (90% of the publications use four or fewer datasets), with an average of 2.7 datasets per publication. Two publications stand out in particular, with 9 datasets (Chakraborty et al. [80]), and 11 datasets (Do et al. [204]) respectively. In accordance with existing work, new publications should evaluate their bias mitigation methods on three datasets, and if possible more.

5.3 Synthetic Data

In addition to the 81 existing datasets for experiments, 54 publications created synthetic datasets to evaluate their bias mitigation method. Moreover, we found 3 publications that use semi-synthetic data (i.e., modify existing datasets to be applicable for evaluating bias mitigation methods) in their experiments [99], [263], [290].

The created datasets range from hundreds of data points [127], [240], [303], [310] to 100,000 and above [43], [134], [179], [186]. While the sampling procedures are well described, some publications do not state the dataset size used for experiments [180], [191], [212], [251], [270], [284], [362], [364].

As exemplary data creation procedure, we briefly outline the data generation approach applied by Zafar et al. [234], as it is the most frequently adapted approach by other publications [83], [180], [220], [239], [285], [296], [309], [328]. In particular, Zafar et al. [234] generated $4,000$ binary class labels. These are augmented with 2-dimensional user features which are drawn from different Gaussian distributions. Lastly, the sensitive attribute is then drawn from a Bernoulli distribution.

5.4 Data-split

In this section we analyze whether existing publications provided information on the data splits, in particular what sizing has been chosen. Moreover, we investigate how often experiments have been repeated with such data splits, to account for training instability [17]. Our focus lies on the data-splits used when evaluating the bias mitigation methods (e.g., we are not interested in data-splits that are applied prior for hyperparameter tuning of classification models [71], [91], [127], [254], [269], [332], [339], [340], [401]).

Among the 324 publications that carry out experiments, 232 provide information on the data-split used and 143 provide information on the number of *runs* (different splits) performed. The high amount of publications that do not provide information on the data-split sizes could be explained by the fact that some of the 81 datasets provided default splits. For example. the Adult dataset has a pre-defined train-test split of 70%-30%, and Cotter et al. [258] used designated data splits for four datasets.

A widely adopted approach for addressing data-splits for applying bias mitigation methods is k-fold cross validation. Such methods divide the data in k partitions and use each part once for testing and the remaining $k-1$ partitions for training. Overall, 47 publication applied cross validation: 10-fold (23 times), 5-fold (21 times), 3-fold (twice), 20-fold (once), and once without specification of k [169] .

If the data-splits are not derived from k-folds, the most popular sizes (i.e., train split size - test split size) are:

- 80%-20% (39 times);
- 70%-30% (35 times);
- 67%-33% (16 times);
- 50%-50% (11 times);
- 60%-40% (5 times);
- 75%-25% (5 times).

In addition to these regular sized datasplits, there are 23 publication which divide the data into very "specific" splits. For example, Quadrianto et al. [123] divided the Adult dataset into $28,222$ training, $15,000$ and $2,000$ validation instance. Another example are Liu and Vicente [296], who chose 5.000 training instances at random, using the remaining $40,222$ instances for testing.

Once the data is split in training and testing data, experiments are repeated 10 times in 54 out of 143 and 5 times in 42 out of 143 cases. The most repetitions are performed by da Cruz [317], who trained $48,000$ models per dataset to evaluate different hyperparameter settings.

We have found 16 publications that use different train and test splits for experiments on multiple datasets. Reasons for that can be found in the stability of bias mitigation methods when dealing with a large amount of training data [166].

While most publications split the data in two parts (i.e., training and test split), there are 36 publication that use validation splits as well. The sizes for validation splits range from 5% to 30%, whereas the most common split uses 60% training data, 20% testing data, and 20% validation data. Furthermore, Mishler and Kennedy [193] allow for a division of the data in up to five different splits for evaluating their ensemble learning procedure.

Bias mitigation methods that process data in a streaming [43], [75], [176], [318], [334], federated learning [73], [87], [150], [288], [321], multi-source [84], sequential [94], [156], [278], [286] fashion need to be addressed differently, as they use small subsets of the training data instead of using all at once.

6 Fairness Metrics

Fairness metrics play an integral part in the bias mitigation process. First they are used to determine the degree of bias a classification model exhibits before applying bias mitigation methods. Afterwards, the effectiveness of bias mitigation methods can be determined by measuring the same metrics after the mitigation procedure.

Recent fairness literature has introduced a variety of different fairness metrics, that each emphasize different aspect of classification performance.

To provide a structured overview of such a large amount of metrics, we devise metric categories, and take into account the classifications by Catan and Haas [24], and Verma and Rubin [21]. Overall we categorize the metrics used in the 341 publications in six categories:

- Definitions based on labels in dataset;
- Definitions based on predicted outcome;
- Definitions based on predicted and actual outcomes;
- Definitions based on predicted probabilities and actual outcome;
- Definitions based on similarity;
- Definitions based on causal reasoning;

In the following, we provide information on how these metric types have been used. In total, we found 111 unique metrics that have been used by the 324 publications that performed experiments. Most publications consider a binary setting (i.e., two populations groups and two class labels for prediction), whereas fairness has also been measured for non-binary sensitive attributes [46], [274], [275], [325], [380], and multi-class predictions [46], [379].

While some of the categories only contain few different metrics (Definitions based on labels in dataset, Definitions Based on Predicted Probabilities and Actual Outcome and Definitions Based on Similarity all have 13 or fewer different metrics); *Definitions Based on Predicted Outcome* have 22, *Definitions Based on Predicted and Actual Outcomes* have 33, and *Definitions Based on Causal Reasoning* 26 different metrics. Therefore, we outline the most frequently used metrics for *Definitions Based on Predicted and Actual Outcomes* and *Definitions Based on Causal Reasoning*.

On average, publications consider two fairness metrics when evaluating bias mitigation methods, with 45% of the publications only using one fairness metric. The most frequently used metrics are outlined in Table 12, while listing at least one metric per category. For detailed explanations of fairness metrics, we refer to and Verma and Rubin [21].

In addition to quantifying the bias according to prediction tasks, we found metrics that determined fairness in accordance with feature usage (e.g., do users think this feature is fair [124]) and quality of representations [115], [119], [125] (see Section 4.1.4).

6.1 Definitions Based on Labels in Dataset

Fairness definition based on the dataset labels, also known as "dataset metrics", are used to determine the degree of bias in an underlying dataset [402]. One purpose of datasets metrics is determine whether there is a balanced representation of privileged and unprivileged groups in the dataset. This is in particular useful for pre-processing bias mitigation methods,

TABLE 12: Popular fairness metrics. At least one metric for each category is provided.

Name	Section	#	Description
Statistical Parity Difference	6.2	137	Difference of positive predictions per group
Equality of Opportunity	6.3	90	Equal TPR per population groups
Disparate Impact, P-rule	6.2	59	Ratio of positive predictions per group
Equalized Odds	6.3	52	Equal TPR and FPR per population groups
False Positive Rate	6.3	38	False positive rate difference per group
Accuracy Rate Difference	6.3	29	Difference of prediction accuracy per group
	...		...
Causal Discrimination	6.5	7	Different predictions for identical individuals except for protected attribute
Mean Difference	6.1	6	Difference of positive labels per group in the datasets
Mutual information	6.6	4	Mutual information between protected attributes and predictions
	...		...
Strong Demographic Disparity	6.4	1	Demographic parity difference over various decision thresholds

as they are able to impact the data distribution of the training dataset.

Most frequently, datasets metrics are used to measure the disparity in positive labels for population groups, such as Mean Difference, slift or elift [352]. Hereby, Mean Difference is the most popular, used in 6 publications.

Another metric based on dataset labels is Balanced Error Rate (BER) [63]. Xu et al. [63] trained an SVM to compare the error rates when predicting protected attributes for both groups.

6.2 Definitions Based on Predicted Outcome

Definitions based on predicted outcome, or "Parity-based" metrics, are used to determine whether different population groups receive the same degree of favour. For this purpose, only the predicted outcome of the classification needs to be known.

The most popular approach for measuring fairness according to predicted outcome is the concept of *Demographic parity*, which states that privileged and unprivileged groups should receive an equal proportion of positive labels. This can be done as by computing their difference (Statistical Parity Difference) or their ratio (Disparate Impact). Similar to Disparate Impact, the p-rule compares two ratios of positive labels ($group_1/group_2$, $group_2/group_1$) and Among those two ratios, the minimum value is chosen. In addition to numeric bias scores, the disparity of group treatment can also be seen visually [48], [70], [239], [259], [269], [351].

If the direction of bias is of no interest (i.e., it is not important which group receives a favourable treatment), then the absolute bias values can be considered [211], [221], [226], [276]. While it is possible to compute fairness metrics based on differences as well as ratios between two groups, both which have been applied in the past, Žliobaite [25] advised against ratios as they are more challenging to interpret.

6.3 Definitions Based on Predicted and Actual Outcomes

Definitions based on predicted and actual outcomes are used to evaluate the prediction performance of privileged and unprivileged groups (e.g., is the classification model more likely to make errors when dealing with unprivileged groups?). Similar to definitions based on predicted outcomes, the rates for privileged and unprivileged groups are compared.

The most popular metric of this type is *Equality of Opportunity* (used 90 times), followed by *Equalized odds* (used 52 times). While *Equality of Opportunity* is satisfied when populations groups have equal TPR, *Equalized odds* is satisfied if population groups have equal TPR and FPR. In addition to evaluating fairness in according to the confusion matrix (FPR - 38 times, TNR - 8 times), the accuracy rate, difference in accuracy for both groups, has been used 29 times. Moreover, conditional TNR and TPR have been evaluated [60], [142].

6.4 Definitions Based on Predicted Probabilities and Actual Outcome

While Section 6.3 detailed metrics based on actual outcomes and predicted labels, this Section outlines metrics that consider predicted probabilities instead.

Jiang et al. [181] proposed strong demographic disparity (SDD) and SPDD, which are parity metrics computed over a variety of thresholds (i.e., prediction tasks apply a threshold of 0.5 by default). Chzhen et al. [266] also varied thresholds, to compute the Kolmogorov-Smirnov distance. Heidari et al. [243] measured fairness based on positive and negative residual differences. Agarwal et al. [262] computed a Bounded Group Loss (BGL) to minimize the worst loss of any group, according to least squares.

6.5 Definitions Based on Similarity

Definitions based on similarity are concerned with the fair treatment individuals. In particular, it is desired that individuals that exhibit a certain degree of similarity receive the same prediction outcome. For this purpose, different similarity measures have been applied. The most popular similarity metric used is *consistency* or *inconsistency* (used in 4 and 1 publications respectively) [110]. *Consistency* compares the prediction of an individual with the k-nearest-neighbors according the input space [110]. Loung et al. [38] also utilized k-nearest-neighbors, to investigate the difference in predictions for different values of k.

Similarities between individuals have been computed according to ℓ_∞-distance [139], and euclidean distance with weights for features [110]. Individuals have also been treated as similar if they have equal labels [165], are equal except for non-sensitive feature or based on predicted label [78]. If similarity of individuals is determined solely by differences in sensitive features, one is speaking of "causal discrimination" [145], [372].[1]

1. Some publications refer to this as "Counterfactual fairness' [196], [218], [346], but we follow the guidelines of Verma and Rubin [21] and treat counterfactual fairness as a Causal metric.

In contrast to determining similarity computationally, Jung et al. [255] allowed stakeholders to judge whether two individuals should receive the same treatment.

Moreover, Ranzato et al. [189] considered four types of similarity relations (NOISE, CAT, NOISE-CAT, CONDITIONAL-ATTRIBUTE), when dealing with numerical and categorical features. Verma et al. [78] considered two types of similarities: input space (identical on non-sensitive features), output space (identical prediction). Lahoti et al. [127] built a similarity graph to detect similar individuals. This graph is built based on pairwise information on individuals that should be treated equally with respect to a given task.

6.6 Causal Reasoning

Fairness definitions based on causal reasoning take causal graphs in account to evaluate relationships between sensitive attributes and outcomes [21].

For example, Counterfactual fairness states that a causal graph is fair, if the prediction does not depend on descendants of the protected attribute [301]. This definition has been adopted by four publications. Moreover, the impact of protected attributes on the decision has been observed in two ways: direct and indirect prejudice [55]. Direct discrimination occurs when the treatment is based on sensitive attributes. Indirect discrimination results in biased decision for population groups based on non-sensitive attributes, which might appear to be neutrals. This could occur due to statistical dependencies between protected and non-protected attributes.

Direct and indirect discrimination can be modelled based on the causal effect along paths taken in causal graphs [55]. To measure indirect discrimination, Prejudice Index (PI) or Normalized Prejudice Index (NPI) haven been applied four times [162]. NPI quantifies the mutual information between protected attributes and predictions. Mutual information has also been used to determine the fairness of representations [122], [125]. Similar to determining the degree of mutual information between sensitive attributes and labels, the ability to predict sensitive information based on representations has been used in eight publications.

7 BENCHMARKING

After establishing on which datasets bias mitigation methods are applied, and which metrics are used to measure their performance (Section 6), we investigate how they have been benchmarked.

Benchmarking is important for ensuring the performance of bias mitigation methods. Nonetheless, we found 15 out of 324 publications that perform experiments but do not compare results with any type of benchmarking. Therefore, the remaining section addresses 308 publications which: 1) perform experiments; 2) apply benchmarking.

7.1 Baseline

To determine whether bias mitigation methods are able to reduce effectively, different types of baselines have been used.

The most general baseline is to compare the fairness achieved by classification models after applying a bias mitigation method with the fairness of a fairness-agnostic Original Model (OM). If a method is not able to exhibit an improved fairness over a fairness agnostic classification model, then it is not applicable for bias mitigation. Given that this is the minimum requirement for bias mitigation methods, it is the most frequently used baseline (used in 254 out of 308 experiments).

Another baseline method is *suppressing*, which performs a naive attempt of mitigating bias by removing the protected attribute from the training data. However, it has been found that solely removing protected attributes does not remove unfairness [7], [37], as the remaining features are often correlated with the protected attribute. To combat this risk, Kamiran et al. [160] suppressed not only the sensitive feature but also the k-most correlated ones. *Suppressing* has been used in 30 out of 308 experiments.

Random baselines constitute more competitive baselines than solely suppressing the protected attribute. Bias mitigation methods that outperform random baselines show that they are not only able to improve fairness but also able to perform better than naive methods. Random baselines have been used in 13 out of 308 experiments.

Moreover, we found four publications that considered a constant classifier for benchmarking (i.e., a classifier that returns the same label for every instance) [122], [180], [267], [337]. This serves as a fairness-aware baseline, as every individual and population group receive the same treatment.

TABLE 13: Benchmarking against bias mitigation method types. For each bias mitigation category (i.e., pre-, in-, or post-processing), we count the type of benchmarking methods.

		#	None	Type Pre	Type In	Type Post
Type	Pre	114	50	55	37	16
	In	184	66	56	108	51
	Post	52	16	17	25	27

7.2 Benchmarking Against Bias Mitigation Methods

In addition to baselines, we investigate how methods are benchmarked against other, existing bias mitigation methods. In particular, we are interested in which methods are popular, how many bias mitigation methods are used for benchmarking, and to what category these methods belong.

At first, we investigate what type of bias mitigation method are considered for benchmarking (e.g., are pre-processing methods more likely to benchmark against other pre-processing methods or in-/post-processing methods). Table 13 illustrates the results. In particular, # shows how many unique publications propose a given type of bias mitigation method (i.e., there are 114 publications with pre-processing methods). For each of these methods we determine whether they benchmark against pre-, in- or post-processing methods. If no benchmarking against other bias mitigation methods is performed, we count this as "None".

We find that pre-processing methods are the most likely to not benchmark against other bias mitigation methods at 44% (50 out of 114). 36% (66 out of 184) of in-processing methods and 31% (16 out of 52) of post-processing methods do not benchmark against other bias mitigation methods.

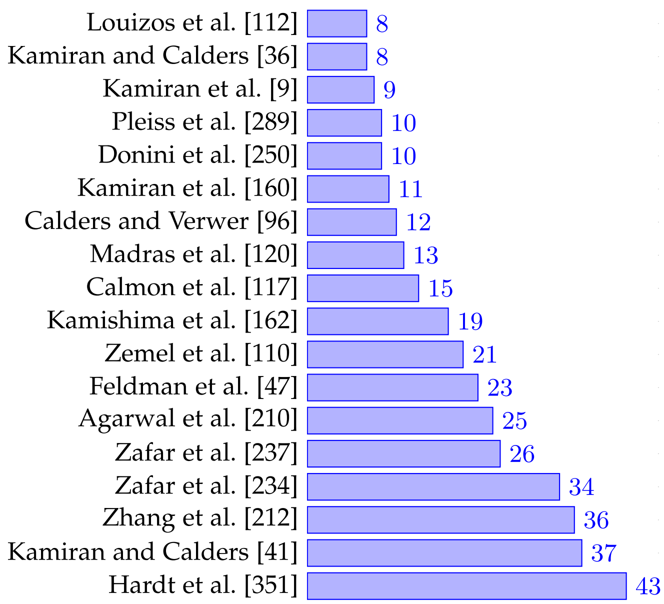


Fig. 4: Most frequently benchmarked publications. For each publication, the number of times it has been used for benchmarking is shown.

Furthermore, we can see that each bias mitigation type is more likely to benchmark against methods of the same type.

In addition to detecting the type of bias mitigation methods for benchmarking, we are interested in what approaches in particular are used for benchmarking. Therefore, we count how often each of the 341 bias mitigation methods we gathered have been used for benchmarking.

Overall, 137 bias mitigation methods have been used as a benchmark by at least one other publication. Figure 4 illustrates the most frequently used bias mitigation methods for benchmarking. Among the 18 listed methods, all of which are used for benchmarking by at least eight other publications, eight are pre-processing, nine in-processing, and four post-processing. Notably, the five most-frequently used methods include each of the three types: sampling and relabelling for pre-processing [41], constraints [234], [237] and adversarial learning [212] for in-processing, and classifier modification for post-processing [351].

7.3 Benchmarking Against Fairness-Unaware Methods

In addition to benchmarking against existing bias mitigation methods, practitioners can use other methods for benchmarking, which are not designed for taking fairness into consideration. Overall, we found 51 publications that use fairness-unaware methods for benchmarking (i.e., using a general data augmentation method to benchmarking fairness-aware resampling).

Table 14 shows the publications that benchmark their proposed method against at least one fairness-unaware methods, according to the type of approach applied. Among the 13 types of approaches, as shown in Section 4.1 - 4.3, seven can be found to benchmark against fairness-unaware methods. This occurs rarely for post-processing methods, six publications in total, with at least one per approach type. 23 and 27 publications for pre-processing and in-processing methods respectively, benchmark against fairness-unaware methods.

TABLE 14: Publications that benchmark against at least one fairness-unaware method.

Type	Category	Section	References
Pre	Sampling	4.1.2	[62], [63], [65], [71], [72], [75], [79], [82], [83], [87], [89]
	Representation	4.1.4	[112], [129], [130], [137], [142], [143] [135], [146], [149], [150], [151], [157]
In	Regularization	4.2.1	[184], [188], [198], [201], [202]
	Constraints	4.2.1	[82], [198], [232], [241], [265], [278], [279], [286]
	Adversarial	4.2.2	[63], [213], [220], [223], [224], [228]
	Adjusted	4.2.4	[79], [184], [286], [298], [318], [335] [75], [325], [326], [330], [340]
Post	Input	4.3.1	[350]
	Classifier	4.3.2	[364], [365]
	Output	4.3.3	[10], [87], [374]

8 Challenges

Research on bias mitigation is fairly young and does therefore enable challenges and opportunities for future research. In this section, we highlight five challenges that we extracted from the collected publications, that call for future action or extension of current work.

8.1 Fairness Definitions

A variety of different metrics have been proposed and used in practice (see Section 6), which can be applied to different use cases. However, with such a variety of metrics it is difficult to evaluate bias mitigation on all and ensure their applicability. Synthesizing or selecting a fixed set of metrics to use is still an open challenge [11], [89], [219], as can be seen by the 111 different fairness metrics obtained in Section 6.

While synthesising existing fairness notions is one problem, it is also relevant to ensure that the used metrics are representative for the problem at hand. Often, this means evaluating fairness in a binary classification problem for two population groups. While this can be the correct way to model fairness scenarios, it is not sufficient to handle all cases, such that future work should focus on multi-class problems [41], [216], [339], [346], [352] and non-binary sensitive attributes, which was mentioned by 15 publications.

Other challenges regarding metrics include the trade-offs when dealing with accuracy and/or multiple fairness metrics [5], [24], [210], [403], as well as the allowance of some degree of discrimination as long it as explainable (e.g., enforcing a fairness criteria completely could lead to unfairness in another) [41], [54], [96], [110].

8.2 Fairness Guarantees

Guarantees are of particular importance when dealing with domains that fall under legislation and regulatory controls [47], [162]. Therefore, it is not always sufficient to establish the effectiveness of a bias mitigation method based on the performance on the test set without any guarantees.

In particular, Dunkelau and Leuschel [18] pointed out that most bias mitigation methods are evaluated on test sets and their applicability to real-world tasks depends on whether the test set reliably represents reality. If that is not the case, fairness guarantees could ensure that bias mitigation methods are able to perform well with regards to unknown data distributions. Therefore, eight publications considered fairness guarantees as a relevant avenue of future work. Similarly, allowing for interpretable and explainable methods can aid in this regard [51], [123], [162], [238].

8.3 Datasets

Another challenge that arises when applying bias mitigation methods is the availability and use of datasets. The most pressing concern is the reliability and access to protected attributes, which was mentioned in nine publications, as this information is often not available in practice [404].

Moreover, it is not guaranteed that the annotation process of the training data is bias free [351]. If possible an unbiased data collection should be enforced [167]. Other options are the debiasing of ground truth labels [85], [145] or use of expert opinions to annotate data [361]. If feasible, more data can be collected [51], [58], which is difficult from a research perspective, as commonly, existing and public datasets are used without the chance to manually collect new samples.

Moreover, the variety of protected attributes addressed in experiments, as found by Kuhlman et al. [16], is lacking diversity, with the majority of cases considering race and gender only. In practice, "collecting more training data" is the most common approach for debiasing, according to interviews conducted by Holstein et al. [404].

8.4 Real-world Applications

While the experiments are conducted on existing, public datasets, it is not clear whether they can be transferred to real-world applications without any adjustments. For example, Hacker and Wiedemann [114] see the challenge of data distributions changing over time, which would require continuous implementations of bias mitigation methods.

Moreover, developers might struggle to detect the relevant population groups to consider when measuring and mitigating bias [404], whereas the datasets investigated in Section 5 often simplify the problem and already provide binarized protected attributes (e.g., in the COMPAS, six "demographic" categories are transformed to "Caucasian" and "not Caucasian" [402]). Therefore, Martinez et al. [319] stated that automatically identifying sub-populations with high-risk during the learning procedure as a field of future work.

Given the multitude of fairness metrics (as seen in Section 6), real world applications could even suffer further unfairness after applying bias mitigation methods due to choosing incorrect criteria [200]. Similarly, showing low bias scores does not necessarily lead to a fair application, as the choice of metrics could be used for "Fairwashing" (i.e., using fake explanations to justify unfair decisions) [364], [405]. Nonetheless, Sylvester and Raff [406] argue that considering fairness criteria while developing ML models is better than considering none, even if the metric is not optimal.

Sharma et al. [66] show the potential of user studies to not only provide bias mitigation methods that work well in a theoretical setting, but to make sure practitioners are willing to use them. In particular, the are interesting in finding how comfortable developers and policy makers are with regards to training data augmentation.

To facilitate the use and implementation of existing bias mitigation methods, metrics and datasets, popular toolkits such as AIF360 [402] and Fairlearn [407] can be used.

8.5 Extension of Experiments

Lastly, a challenge and field of future research is the extension of conducted experiments to allow for more meaningful results.

The most frequently discussed aspect of extending experiments is the consideration of further metrics (in 40 publications). Moreover, the usefulness of bias mitigation methods can be investigated when applied to additional classification models. This was pointed out by 12 publications. Given the 81 datasets that were used at least once, and on average 2.7 datasets used per publication, only eight publications see the consideration of further datasets as a useful consideration for extending their experiments [56], [67], [71], [311], [317], [322], [327], [340].

While the consideration of additional metrics, classification models and datasets does not lead to changes in the training procedure and experimental design, there are also intentions to apply bias mitigation methods to other tasks and contexts, such as recommendations [190], [234], ranking [162], [175], [234] and clustering [162].

9 Conclusion

In this literature survey, we have focused on the adoption of bias mitigation methods to achieve fairness in classification problems and provided an overview of 341 publications. Our survey first categories bias mitigation methods according to their type (i.e., pre-processing, in-processing, post-processing) and illustrates their procedures. We found 123 pre-processing, 212 in-processing, and 56 post-processing methods, showing that in-processing methods are the most commonly used. We devised 13 categories for the three method types, based on their approach (e.g., pre-processing methods can perform sampling). The most frequently applied approaches perform changes to the loss function in an in-processing stage (51 publications applying regularization and 74 applying constraints). Other approaches are less frequently used, with input correction in a post-processing stage only being used twice.

We further provided insights on the evaluation of bias mitigation methods according to three aspects: datasets, metrics, and benchmarking. We found a total of 81 datasets that have been used at least once by one of the 341 publications, among which the Adult dataset is the most popular (used by 77% of publications). Even though 81 datasets are available for evaluating bias mitigation methods, only 2.7 datasets are considered on average.

Similarly, we found a large number of fairness metrics that have been used at least once (111 unique metrics), which we divide in six categories. The most frequently used metrics belong to two categories: 1) Definitions based on predicted outcome; 2) Definitions based on predicted and actual outcomes.

When it comes to benchmarking bias mitigation methods, they can be compared against baselines, other bias mitigation methods, or non-bias mitigation approaches. Among the three baselines we found (original model, suppressing, random), the 82% of bias mitigation methods consider the original model (i.e., the classification model without any bias mitigation applied) as a baseline. Commonly, methods are compared against other bias mitigation methods. 51 publications benchmark against fairness-unaware methods.

Lastly, we list avenues of future work and challenges that have been discerned in the collected publications. This includes the synthesizing of fairness metrics, as there is no consensus reached on what metrics to use. In addition to measuring improvements, future bias mitigation methods can take fairness guarantees in account. The application of bias mitigation methods in practice is challenging, as developers might not be able to detect relevant population groups for which to measure bias and reliability of datasets (i.e., are prior observations biased?). Therefore, we hope that this survey helps researchers and practitioners to gain an understanding of the current, existing bias mitigation approaches and aspects support their development of new methods.

Acknowledgments

This work is supported by the ERC Advanced fellowship grant no. 741278 (EPIC). We would like to thank the members of the community who kindly provided comments and feedback on an earlier version of this paper.

References

[1] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias. propublica," *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

[3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, p. 0049124118782533, 2018.

[4] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, "Learning gender-neutral word embeddings," *EMNLP*, pp. 4847–4853, 2018.

[5] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.

[6] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[7] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568.

[8] B. van Giffen, D. Herhausen, and T. Fahse, "Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods," *Journal of Business Research*, vol. 144, pp. 93–106, 2022.

[9] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 924–929.

[10] F. Kamiran, S. Mansha, A. Karim, and X. Zhang, "Exploiting reject option in classification for social discrimination control," *Information Sciences*, vol. 425, pp. 18–33, 2018.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[12] J. Dastin. (2018, oct) Amazon scraps secret ai recruiting tool that showed bias against women. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[13] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from berkeley," *Science*, vol. 187, no. 4175, pp. 398–404, 1975.

[14] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 339–348.

[15] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," 2011.

[16] C. Kuhlman, L. Jackson, and R. Chunara, "No computation without representation: Avoiding data and algorithm biases through diversity," *arXiv preprint arXiv:2002.11836*, 2020.

[17] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 329–338.

[18] J. Dunkelau and M. Leuschel, "Fairness-aware machine learning," 2019.

[19] M. Hort, J. Zhang, F. Sarro, and M. Harman, "Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021.

[20] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "A comprehensive empirical study of bias mitigation methods for software fairness," *arXiv preprint arXiv:2207.03277*, 2022.

[21] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018, pp. 1–7.

[22] "On-line appendix: Survey results," 2022. [Online]. Available: https://docs.google.com/spreadsheets/d/1kOmbKLMiFgHRSXvgM-O8OW4YKeDIGN0cPPeCGQOMnnA/edit?usp=sharing

[23] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[24] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[25] I. Žliobaite, "A survey on measuring indirect discrimination in machine learning," *arXiv preprint arXiv:1511.00148*, 2015.

[26] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1452, 2022.

[27] X. Zhang and M. Liu, "Fairness in learning-based sequential decision algorithms: A survey," in *Handbook of Reinforcement Learning and Control*. Springer, 2021, pp. 525–555.

[28] W. Zhang, J. C. Weiss, S. Zhou, and T. Walsh, "Fairness amidst non-iid graph data: A literature review," *arXiv preprint arXiv:2202.07170*, 2022.

[29] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, 2020.

[30] E. Soremekun, M. Papadakis, M. Cordy, and Y. L. Traon, "Software fairness: An analysis and survey," *arXiv preprint arXiv:2205.08809*, 2022.

[31] Z. Chen, J. M. Zhang, M. Hort, F. Sarro, and M. Harman, "Fairness testing: A comprehensive survey and analysis of trends," *arXiv e-prints*, pp. arXiv–2207, 2022.

[32] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, "Mitigating gender bias in natural language processing: Literature review," *arXiv preprint arXiv:1906.08976*, 2019.

[33] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of" bias" in nlp," *arXiv preprint arXiv:2005.14050*, 2020.

[34] W. Martin, F. Sarro, Y. Jia, Y. Zhang, and M. Harman, "A survey of app store analysis for software engineering," *IEEE transactions on software engineering*, vol. 43, no. 9, pp. 817–847, 2016.

[35] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings*

of the 18th international conference on evaluation and assessment in software engineering, 2014, pp. 1–10.

[36] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd international conference on computer, control and communication*. IEEE, 2009, pp. 1–6.

[37] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009, pp. 13–18.

[38] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 502–510.

[39] I. Žliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 992–1001.

[40] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 7, pp. 1445–1459, 2012.

[41] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[42] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in prediction," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, p. 3097–3103.

[43] V. Iosifidis, T. N. H. Tran, and E. Ntoutsi, "Fairness-enhancing interventions in stream classification," in *International Conference on Database and Expert Systems Applications*. Springer, 2019, pp. 261–276.

[44] H. Sun, K. Wu, T. Wang, and W. H. Wang, "Towards fair and robust classification," in *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2022, pp. 356–376.

[45] E. Seker, J. R. Talburt, and M. L. Greer, "Preprocessing to address bias in healthcare data," *Studies in Health Technology and Informatics*, vol. 294, pp. 327–331, 2022.

[46] I. Alabdulmohsin, J. Schrouff, and O. Koyejo, "A reduction to binary approach for debiasing multiclass datasets," *arXiv preprint arXiv:2205.15860*, 2022.

[47] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[48] K. Lum and J. Johndrow, "A statistical framework for fair predictive algorithms," *arXiv preprint arXiv:1610.08077*, 2016.

[49] H. Wang, B. Ustun, F. P. Calmon, and S. Harvard, "Avoiding disparate impact with counterfactual distributions," in *NeurIPS Workshop on Ethical, Social and Governance Issues in AI*, 2018.

[50] H. Wang, B. Ustun, and F. Calmon, "Repairing without retraining: Avoiding disparate impact with counterfactual distributions," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6618–6627.

[51] J. E. Johndrow and K. Lum, "An algorithm for removing sensitive information: application to race-independent recidivism prediction," *The Annals of Applied Statistics*, vol. 13, no. 1, pp. 189–220, 2019.

[52] T. Li, Z. Tang, T. Lu, and X. M. Zhang, "'propose and review': Interactive bias mitigation for machine classifiers," *Available at SSRN 4139244*, 2022.

[53] Y. Li, L. Meng, L. Chen, L. Yu, D. Wu, Y. Zhou, and B. Xu, "Training data debugging for the fairness of machine learning software," in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022, pp. 2215–2227.

[54] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling," in *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Citeseer, 2010, pp. 1–6.

[55] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3929–3935.

[56] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 853–862.

[57] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 570–575.

[58] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[59] V. Iosifidis and E. Ntoutsi, "Dealing with bias via data augmentation in supervised learning scenarios," *Jo Bates Paul D. Clough Robert Jäschke*, vol. 24, 2018.

[60] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Interventional fairness: Causal database repair for algorithmic fairness," in *Proceedings of the 2019 International Conference on Management of Data*, 2019, pp. 793–810.

[61] V. Zelaya, P. Missier, and D. Prangle, "Parametrised data sampling for fairness optimisation," *KDD XAI*, 2019.

[62] D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu, "Achieving causal fairness through generative adversarial networks," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[63] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan+: Achieving fair data generation and classification through generative adversarial nets," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 1401–1406.

[64] V. Iosifidis, B. Fetahu, and E. Ntoutsi, "Fae: A fairness-aware ensemble framework," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 1375–1380.

[65] A. Abusitta, E. Aïmeur, and O. A. Wahab, "Generative adversarial networks for mitigating biases in machine learning systems," *arXiv preprint arXiv:1905.09972*, 2019.

[66] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, "Data augmentation for discrimination prevention and bias disambiguation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 358–364.

[67] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies, *Fairway: A Way to Build Fair ML Software*. New York, NY, USA: Association for Computing Machinery, 2020, p. 654–665. [Online]. Available: https://doi.org/10.1145/3368089.3409697

[68] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 702–712.

[69] T. Hu, V. Iosifidis, W. Liao, H. Zhang, M. Y. Yang, E. Ntoutsi, and B. Rosenhahn, "Fairnn-conjoint learning of fair representations for fair decisions," in *International Conference on Discovery Science*. Springer, 2020, pp. 581–595.

[70] A. Morano, "Bias mitigation for automated decision making systems," Ph.D. dissertation, Politecnico di Torino, 2020.

[71] S. Yan, H.-t. Kao, and E. Ferrara, "Fair class balancing: enhancing model fairness without observing sensitive attributes," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1715–1724.

[72] L. E. Celis, V. Keswani, and N. Vishnoi, "Data preprocessing to mitigate bias: A maximum entropy based approach," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1349–1359.

[73] A. Abay, Y. Zhou, N. Baracaldo, S. Rajamoni, E. Chuba, and H. Ludwig, "Mitigating bias in federated learning," *arXiv preprint arXiv:2012.02447*, 2020.

[74] T. Salazar, M. S. Santos, H. Araújo, and P. H. Abreu, "Fawos: Fairness-aware oversampling algorithm based on distributions of sensitive attributes," *IEEE Access*, 2021.

[75] W. Zhang, A. Bifet, X. Zhang, J. C. Weiss, and W. Nejdl, "Farf: A fair and adaptive random forests classifier," in *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part II*, 2021, pp. 245–256.

[76] C.-Y. Chuang and Y. Mroueh, "Fair mixup: Fairness via interpolation," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=DNl5s5BXeBn

[77] J. J. Amend and S. Spurlock, "Improving machine learning fairness with sampling and adversarial learning," *Journal of Computing Sciences in Colleges*, vol. 36, no. 5, pp. 14–23, 2021.

[78] S. Verma, M. Ernst, and R. Just, "Removing biased data to improve fairness and accuracy," *arXiv preprint arXiv:2102.03054*, 2021.

[79] A. F. Cruz, P. Saleiro, C. Belém, C. Soares, and P. Bizarro, "Promoting fairness through hyperparameter optimization," in *2021 IEEE International Conference on Data Mining (ICDM)*, 2021, pp. 1036–1041.

[80] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? how? what to do?" in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 429–440. [Online]. Available: https://doi.org/10.1145/3468264.3468537

[81] T. Jang, F. Zheng, and X. Wang, "Constructing a fair classifier with generated fair data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7908–7916.

[82] W. Du and X. Wu, "Fair and robust classification under sample selection bias," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2999–3003. [Online]. Available: https://doi.org/10.1145/3459637.3482104

[83] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Sample selection for fair and robust training," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[84] E. Iofinova, N. Konstantinov, and C. H. Lampert, "Flea: Provably fair multisource learning from unreliable training data," *arXiv preprint arXiv:2106.11732*, 2021.

[85] Z. Yu, "Fair balance: Mitigating machine learning bias against multiple protected attributes with data balancing," *CoRR*, vol. abs/2107.08310, 2021. [Online]. Available: https://arxiv.org/abs/2107.08310

[86] A. Singh, J. Singh, A. Khan, and A. Gupta, "Developing a novel fair-loan classifier through a multi-sensitive debiasing pipeline: Dualfair," *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 240–253, 2022.

[87] S. Pentyala, N. Neophytou, A. Nascimento, M. De Cock, and G. Farnadi, "Privfairfl: Privacy-preserving group fairness in federated learning," *arXiv preprint arXiv:2205.11584*, 2022.

[88] A. Rajabi and O. O. Garibay, "Tabfairgan: Fair tabular data generation with generative adversarial networks," *Machine Learning and Knowledge Extraction*, vol. 4, no. 2, pp. 488–501, 2022.

[89] D. Dablain, B. Krawczyk, and N. Chawla, "Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning," *arXiv preprint arXiv:2207.06084*, 2022.

[90] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Maat: A novel ensemble approach to addressing fairness and performance bugs for machine learning software," in *Proceedings of the 2022 ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE'22*, 2022.

[91] P. Li and H. Liu, "Achieving fairness at no utility cost via data reweighing with influence," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 917–12 930.

[92] J. Chakraborty, S. Majumder, and H. Tu, "Fair-ssl: Building fair ml software with less data," in *International Workshop on Equitable Data and Technology (FairWare '22)*, 2022.

[93] J. Wang, X. E. Wang, and Y. Liu, "Understanding instance-level impact of fairness constraints," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 23 114–23 130. [Online]. Available: https://proceedings.mlr.press/v162/wang22ac.html

[94] A. A. Almuzaini, C. A. Bhatt, D. M. Pennock, and V. K. Singh, "Abcinml: Anticipatory bias correction in machine learning applications," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1552–1560. [Online]. Available: https://doi.org/10.1145/3531146.3533211

[95] J. Chai and X. Wang, "Fairness with adaptive weights," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2853–2866.

[96] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[97] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 656–666.

[98] M. Gupta, A. Cotter, M. M. Fard, and S. Wang, "Proxy fairness," *arXiv preprint arXiv:1806.11212*, 2018.

[99] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Fairness through causal awareness: Learning causal latent-variable models for biased data," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 349–358.

[100] L. Oneto, M. Doninini, A. Elders, and M. Pontil, "Taking advantage of multitask learning for fair classification," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 227–237.

[101] D. Wei, K. N. Ramamurthy, and F. d. P. Calmon, "Optimized score transformation for fair classification," *Proceedings of Machine Learning Research*, vol. 108, 2020.

[102] T. Kehrenberg, Z. Chen, and N. Quadrianto, "Tuning fairness by balancing target labels," *Frontiers in artificial intelligence*, vol. 3, p. 33, 2020.

[103] V. Grari, S. Lamprier, and M. Detyniecki, "Fairness without the sensitive attribute via causal variational autoencoder," *arXiv preprint arXiv:2109.04999*, 2021.

[104] C. Chen, Y. Liang, X. Xu, S. Xie, Y. Hong, and K. Shu, "On fair classification with mostly private sensitive attributes," *arXiv preprint arXiv:2207.08336*, 2022.

[105] Y. Liang, C. Chen, T. Tian, and K. Shu, "Joint adversarial learning for cross-domain fair classification," *arXiv preprint arXiv:2206.03656*, 2022.

[106] S. Jung, S. Chun, and T. Moon, "Learning fair classifiers with partially annotated group labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 348–10 357.

[107] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, A. Roth, and S. Sharifi-Malvajerdi, "Multiaccurate proxies for downstream fairness," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1207–1239. [Online]. Available: https://doi.org/10.1145/3531146.3533180

[108] S. Wu, M. Gong, B. Han, Y. Liu, and T. Liu, "Fair classification with instance-dependent label noise," in *Conference on Causal Learning and Reasoning*. PMLR, 2022, pp. 927–943.

[109] V. M. Suriyakumar, M. Ghassemi, and B. Ustun, "When personalization harms: Reconsidering the use of group attributes in prediction," *arXiv preprint arXiv:2206.02058*, 2022.

[110] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[111] H. Edwards and A. Storkey, "Censoring representations with an adversary," *arXiv preprint arXiv:1511.05897*, 2015.

[112] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel, "The variational fair autoencoder," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.00830

[113] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, "Controllable invariance through adversarial feature learning," *Advances in neural information processing systems*, vol. 30, 2017.

[114] P. Hacker and E. Wiedemann, "A continuous framework for fairness," *arXiv preprint arXiv:1712.07924*, 2017.

[115] D. McNamara, C. S. Ong, and R. C. Williamson, "Provably fair representations," *arXiv preprint arXiv:1710.04394*, 2017.

[116] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair kernel learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 339–355.

[117] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[118] J. Komiyama and H. Shimao, "Two-stage algorithm for fairness-aware machine learning," *arXiv preprint arXiv:1710.04924*, 2017.

[119] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala, "The price of fair pca: One extra dimension," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 976–10 987.

[120] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3384–3393.

[121] F. du Pin Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis,"

IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 5, pp. 1106–1119, 2018.

[122] D. Moyer, S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan, "Invariant representations without adversarial training," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 9102–9111.

[123] N. Quadrianto, V. Sharmanska, and O. Thomas, "Neural styling for interpretable fair representations," *arXiv preprint arXiv:1810.06755*, 2018.

[124] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[125] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, "Learning controllable fair representations," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2164–2173.

[126] X. Wang and H. Huang, "Approaching machine learning fairness through adversarial network," *arXiv preprint arXiv:1909.03013*, 2019.

[127] P. Lahoti, K. P. Gummadi, and G. Weikum, "Operationalizing individual fairness with pairwise fair representations," *Proc. VLDB Endow.*, vol. 13, no. 4, p. 506–518, dec 2019. [Online]. Available: https://doi.org/10.14778/3372716.3372723

[128] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, and C. Wang, "Learning fair representations via an adversarial framework," *arXiv preprint arXiv:1904.13341*, 2019.

[129] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *2019 ieee 35th international conference on data engineering (icde)*. IEEE, 2019, pp. 1334–1345.

[130] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *International conference on machine learning*. PMLR, 2019, pp. 1436–1445.

[131] P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes, "Obtaining fairness using optimal transport theory," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2357–2365.

[132] N. Quadrianto, V. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8227–8236.

[133] H. Zhao, A. Coston, T. Adel, and G. J. Gordon, "Conditional learning of fair representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=Hkekl0NFPr

[134] M. Zehlike, P. Hacker, and E. Wiedemann, "Matching code and law: achieving algorithmic fairness with optimal transport," *Data Mining and Knowledge Discovery*, vol. 34, no. 1, pp. 163–200, 2020.

[135] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "Fairness by learning orthogonal disentangled representations," in *European Conference on Computer Vision*. Springer, 2020, pp. 746–761.

[136] Z. Tan, S. Yeom, M. Fredrikson, and A. Talwalkar, "Learning fair representations for kernel models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 155–166.

[137] A. Jaiswal, D. Moyer, G. Ver Steeg, W. AbdAlmageed, and P. Natarajan, "Invariant representations through adversarial forgetting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4272–4279.

[138] R. Madhavan and M. Wadhwa, "Fairness-aware learning with prejudice free representations," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2137–2140.

[139] A. Ruoss, M. Balunovic, M. Fischer, and M. Vechev, "Learning certified individually fair representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7584–7596, 2020.

[140] J.-Y. Kim and S.-B. Cho, "Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck." in *SafeAI@ AAAI*, 2020, pp. 105–112.

[141] H. Fong, V. Kumar, A. Mehrotra, and N. K. Vishnoi, "Fairness for auc via feature augmentation," *arXiv preprint arXiv:2111.12823*, 2021.

[142] R. Salazar, F. Neutatz, and Z. Abedjan, "Automated feature engineering for algorithmic fairness," *Proceedings of the VLDB Endowment*, vol. 14, no. 9, pp. 1694–1702, 2021.

[143] U. Gupta, A. Ferber, B. Dilkina, and G. Ver Steeg, "Controllable guarantees for fair outcomes via contrastive information estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7610–7619.

[144] V. Grari, O. E. Hajouji, S. Lamprier, and M. Detyniecki, "Learning unbiased representations via rényi minimization," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 749–764.

[145] W. Zhu, H. Zheng, H. Liao, W. Li, and J. Luo, "Learning bias-invariant representation by cross-sample mutual information minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 002–15 012.

[146] C. Oh, H. Won, J. So, T. Kim, Y. Kim, H. Choi, and K. Song, "Learning fair representation via distributional contrastive disentanglement," *arXiv preprint arXiv:2206.08743*, 2022.

[147] S. Agarwal and A. Deshpande, "On the power of randomization in fair classification and representation," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1542–1551. [Online]. Available: https://doi.org/10.1145/3531146.3533209

[148] C. Wu, F. Wu, T. Qi, and Y. Huang, "Semi-fairvae: Semi-supervised fair representation learning with adversarial variational autoencoder," *arXiv preprint arXiv:2204.00536*, 2022.

[149] C. Shui, Q. Chen, J. Li, B. Wang, and C. Gagné, "Fair representation learning through implicit path alignment," *arXiv preprint arXiv:2205.13316*, 2022.

[150] T. Qi, F. Wu, C. Wu, L. Lyu, T. Xu, Z. Yang, Y. Huang, and X. Xie, "Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning," *arXiv preprint arXiv:2206.03200*, 2022.

[151] M. Balunović, A. Ruoss, and M. Vechev, "Fair normalizing flows," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=BrFIKuxrZE

[152] P. Kairouz, J. Liao, C. Huang, M. Vyas, M. Welfert, and L. Sankar, "Generating fair universal representations using adversarial models," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1970–1985, 2022.

[153] S. Liu, S. Sun, and J. Zhao, "Fair transfer learning with factor variational auto-encoder," *Neural Processing Letters*, pp. 1–13, 2022.

[154] M. Cerrato, A. V. Coronel, M. Köppel, A. Segner, R. Esposito, and S. Kramer, "Fair interpretable representation learning with correction vectors," *arXiv preprint arXiv:2202.03078*, 2022.

[155] M. M. Kamani, F. Haddadpour, R. Forsati, and M. Mahdavi, "Efficient fair principal component analysis," *Machine Learning*, pp. 1–32, 2022.

[156] M. Rateike, A. Majumdar, O. Mineeva, K. P. Gummadi, and I. Valera, "Don't throw it away! the utility of unlabeled data in fair decision making," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1421–1433.

[157] S. Galhotra, K. Shanmugam, P. Sattigeri, and K. R. Varshney, "Causal feature selection for algorithmic fairness," in *Proceedings of the 2022 International Conference on Management of Data*, ser. SIGMOD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 276–285. [Online]. Available: https://doi.org/10.1145/3514221.3517909

[158] J.-Y. Kim and S.-B. Cho, "An information theoretic approach to reducing algorithmic bias for machine learning," *Neurocomputing*, vol. 500, pp. 26–38, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222005987

[159] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[160] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 869–874.

[161] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 643–650.

[162] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 35–50.

[163] G. Ristanoski, W. Liu, and J. Bailey, "Discrimination aware classification for imbalanced datasets," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 1529–1532.

[164] B. Fish, J. Kun, and A. D. Lelkes, "Fair boosting: a case study," in *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Citeseer, 2015.
[165] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*, 2017.
[166] Y. Bechavod and K. Ligett, "Penalizing unfairness in binary classification," *arXiv preprint arXiv:1707.00044*, 2017.
[167] N. Quadrianto and V. Sharmanska, "Recycling privileged learning and distribution matching for fairness," *Advances in Neural Information Processing Systems*, vol. 30, pp. 677–688, 2017.
[168] E. Raff, J. Sylvester, and S. Mills, "Fair forests: Regularized tree induction to minimize model bias," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 243–250.
[169] N. Goel, M. Yaghini, and B. Faltings, "Non-discriminatory machine learning through convex fairness criteria," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[170] S. A. Enni and I. Assent, "Using balancing terms to avoid discrimination in classification," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 947–952.
[171] J. Mary, C. Calauzenes, and N. El Karoui, "Fairness-aware learning for continuous attributes and treatments," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4382–4391.
[172] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, "Putting fairness principles into practice: Challenges, metrics, and improvements," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 453–459.
[173] W. Zhang, X. Tang, and J. Wang, "On fairness-aware learning for non-discriminative decision-making," in *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2019, pp. 1072–1079.
[174] S. Aghaei, M. J. Azizi, and P. Vayanos, "Learning optimal and fair decision trees for non-discriminative decision-making," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1418–1426.
[175] L. Huang and N. Vishnoi, "Stable and fair classification," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2879–2890.
[176] W. Zhang and E. Ntoutsi, "Faht: an adaptive fairness-aware decision tree classifier," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 1480–1486.
[177] M. Tavakol, "Fair classification with counterfactual learning," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2073–2076.
[178] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn, "R\'enyi fair inference," in *8th International Conference on Learning Representations, ICLR 2020*, 2020.
[179] P. G. Di Stefano, J. M. Hickey, and V. Vasileiou, "Counterfactual fairness: removing direct effects through regularization," *arXiv preprint arXiv:2002.10774*, 2020.
[180] J. S. Kim, J. Chen, and A. Talwalkar, "Fact: A diagnostic for group fairness trade-offs," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5264–5274.
[181] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 862–872.
[182] Y. Romano, S. Bates, and E. Candes, "Achieving equalized odds by resampling sensitive attributes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 361–371, 2020.
[183] S. Ravichandran, D. Khurana, B. Venkatesh, and N. U. Edakunni, "Fairxgboost: Fairness-aware classification in xgboost," *arXiv preprint arXiv:2009.01442*, 2020.
[184] W. Liu, X. Wang, X. Lu, J. Cheng, B. Jin, X. Wang, and H. Zha, "Fair differential privacy can mitigate the disparate impact on model accuracy," 2021. [Online]. Available: https://openreview.net/forum?id=IqVB8e0DlUd
[185] K. N. Keya, R. Islam, S. Pan, I. Stockwell, and J. R. Foulds, "Equitable allocation of healthcare resources with fair cox models," *arXiv preprint arXiv:2010.06820*, 2020.
[186] J. M. Hickey, P. G. D. Stefano, and V. Vasileiou, "Fairness by explicability and adversarial shap learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020, pp. 174–190.
[187] M. M. Kamani, "Multiobjective optimization approaches for bias mitigation in machine learning," 2020.
[188] W. Zhang and J. C. Weiss, "Fair decision-making under uncertainty," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 886–895.
[189] F. Ranzato, C. Urban, and M. Zanella, "Fairness-aware training of decision trees by abstract interpretation," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1508–1517.
[190] J. Kang, T. Xie, X. Wu, R. Maciejewski, and H. Tong, "Multifair: Multi-group fairness in machine learning," *arXiv preprint arXiv:2105.11069*, 2021.
[191] V. Grari, S. Lamprier, and M. Detyniecki, "Fairness-aware neural rényi minimization for continuous features," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 2262–2268.
[192] Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi, "Understanding and improving fairness-accuracy trade-offs in multi-task learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1748–1757.
[193] A. Mishler and E. Kennedy, "Fade: Fair double ensemble learning for observable and counterfactual outcomes," *arXiv preprint arXiv:2109.00173*, 2021.
[194] A. Lowy, R. Pavan, S. Baharlouei, M. Razaviyayn, and A. Beirami, "Fermi: Fair empirical risk minimization via exponential r\'enyi mutual information," *arXiv preprint arXiv:2102.12586*, 2021.
[195] T. Zhao, E. Dai, K. Shu, and S. Wang, "You can still achieve fairness without sensitive attributes: Exploring biases in non-sensitive features," *arXiv preprint arXiv:2104.14537*, 2021.
[196] M. Yurochkin and Y. Sun, "Sensei: Sensitive set invariance for enforcing individual fairness," in *International Conference on Learning Representations*, 2021.
[197] T. Zhao, E. Dai, K. Shu, and S. Wang, "Towards fair classifiers without sensitive attributes: Exploring biases in related features," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 1433–1442.
[198] J. Wang, Y. Li, and C. Wang, "Synthesizing fair decision trees via iterative constraint solving," in *International Conference on Computer Aided Verification*. Springer, 2022, pp. 364–385.
[199] Z. Deng, J. Zhang, L. Zhang, T. Ye, Y. Coley, W. J. Su, and J. Zou, "Fifa: Making fairness more generalizable in classifiers trained on imbalanced data," *arXiv preprint arXiv:2206.02792*, 2022.
[200] J. Lee, Y. Bu, P. Sattigeri, R. Panda, G. W. Wornell, L. Karlinsky, and R. Schmidt Feris, "A maximal correlation framework for fair machine learning," *Entropy*, vol. 24, no. 4, p. 461, 2022.
[201] W. Zhang and J. C. Weiss, "Longitudinal fairness with censorship," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 235–12 243.
[202] Z. Jiang, X. Han, C. Fan, F. Yang, A. Mostafavi, and X. Hu, "Generalized demographic parity for group fairness," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=YigKlMJwjye
[203] J. Lee, Y. Bu, P. Sattigeri, R. Panda, G. Wornell, L. Karlinsky, and R. Feris, "A maximal correlation approach to imposing fairness in machine learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3523–3527.
[204] H. Do, P. Putzel, A. S. Martin, P. Smyth, and J. Zhong, "Fair generalized linear models with a convex penalty," in *International Conference on Machine Learning*. PMLR, 2022, pp. 5286–5308.
[205] P. Patil and K. Purcell, "Decorrelation-based deep learning for bias mitigation," *Future Internet*, vol. 14, no. 4, p. 110, 2022.
[206] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," *arXiv preprint arXiv:1707.00075*, 2017.
[207] S. Gillen, C. Jung, M. Kearns, and A. Roth, "Online learning with an unknown fairness metric," *Advances in neural information processing systems*, vol. 31, 2018.
[208] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 2564–2572. [Online]. Available: http://proceedings.mlr.press/v80/kearns18a.html
[209] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: an application to recidivism prediction," *arXiv preprint arXiv:1807.00199*, 2018.

[210] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[211] E. Raff and J. Sylvester, "Gradient reversal against discrimination: A fair neural network learning approach," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2018, pp. 189–198.

[212] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018, pp. 335–340.

[213] B. Sadeghi, R. Yu, and V. Boddeti, "On the global optima of kernelized adversarial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7971–7979.

[214] T. Adel, I. Valera, Z. Ghahramani, and A. Weller, "One-network adversarial fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2412–2420.

[215] H. Zhao and G. Gordon, "Inherent tradeoffs in learning fair representations," *Advances in neural information processing systems*, vol. 32, 2019.

[216] L. E. Celis and V. Keswani, "Improved adversarial learning for fair classification," *arXiv preprint arXiv:1901.10443*, 2019.

[217] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki, "Fair adversarial gradient tree boosting," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1060–1065.

[218] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ml models with sensitive subspace robustness," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1gdkxHFDH

[219] A. S. Garcia de Alford, S. K. Hayden, N. Wittlin, and A. Atwood, "Reducing age bias in machine learning: An algorithmic approach," *SMU Data Science Review*, vol. 3, no. 2, p. 11, 2020.

[220] Y. Roh, K. Lee, S. Whang, and C. Suh, "Fr-train: A mutual information-based approach to fair and robust training," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8147–8157.

[221] P. Delobelle, P. Temple, G. Perrouin, B. Frénay, P. Heymans, and B. Berendt, "Ethical adversaries: Towards mitigating unfairness with adversarial machine learning," in *Bias and Fairness in AI (BIAS 2020)*, 2020.

[222] A. Rezaei, R. Fathony, O. Memarrast, and B. Ziebart, "Fairness for robust log loss classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5511–5518.

[223] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," *Advances in neural information processing systems*, vol. 33, pp. 728–740, 2020.

[224] A. Rezaei, A. Liu, O. Memarrast, and B. D. Ziebart, "Robust fairness under covariate shift," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9419–9427.

[225] G. Tao, W. Sun, T. Han, C. Fang, and X. Zhang, "Ruler: Discriminative and iterative adversarial training for deep neural network fairness," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022*, 2022.

[226] A. Petrović, M. Nikolić, S. Radovanović, B. Delibašić, and M. Jovanović, "Fair: Fair adversarial instance re-weighting," *Neurocomputing*, vol. 476, pp. 14–37, 2022.

[227] J. Yang, A. A. Soltan, Y. Yang, and D. A. Clifton, "Algorithmic fairness and bias mitigation for clinical machine learning: Insights from rapid covid-19 diagnosis by adversarial learning," *medRxiv*, 2022.

[228] M. Yazdani-Jahromi, A. Rajabi, A. Tayebi, and O. O. Garibay, "Distraction is all you need for fairness," *arXiv preprint arXiv:2203.07593*, 2022.

[229] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[230] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," in *2013 IEEE 13th international conference on data mining*. IEEE, 2013, pp. 71–80.

[231] K. Fukuchi and J. Sakuma, "Fairness-aware learning with restriction of universal dependency using f-divergences," *arXiv preprint arXiv:1506.07721*, 2015.

[232] K. Fukuchi, T. Kamishima, and J. Sakuma, "Prediction with model-based neutrality," *IEICE TRANSACTIONS on Information and Systems*, vol. 98, no. 8, pp. 1503–1516, 2015.

[233] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander, "Satisfying real-world goals with dataset constraints," in *Advances in Neural Information Processing Systems*, 2016, pp. 2415–2423.

[234] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, 2017, pp. 962–970.

[235] C. Russell, M. Kusner, C. Loftus, and R. Silva, "When worlds collide: integrating different counterfactual assumptions in fairness," in *Advances in neural information processing systems*, vol. 30. NIPS Proceedings, 2017.

[236] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[237] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[238] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Conference on Learning Theory*. PMLR, 2017, pp. 1920–1953.

[239] M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi, and A. Weller, "From parity to preference-based notions of fairness in classification," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 228–238.

[240] M. Olfat and A. Aswani, "Spectral algorithms for computing fair support vector machines," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1933–1942.

[241] H. Narasimhan, "Learning with complex loss functions and constraints," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1646–1654.

[242] J. Zhang and E. Bareinboim, "Fairness in decision-making—the causal explanation formula," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[243] H. Heidari, C. Ferrari, K. Gummadi, and A. Krause, "Fairness behind a veil of ignorance: A welfare analysis for automated decision making," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[244] M. Kim, O. Reingold, and G. Rothblum, "Fairness through computationally-bounded awareness," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[245] G. Farnadi, B. Babaki, and L. Getoor, "Fairness in relational domains," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 108–114.

[246] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[247] Y. Wu, L. Zhang, and X. Wu, "Fairness-aware classification: Criterion, convexity, and bounds," *arXiv preprint arXiv:1809.04737*, 2018.

[248] J. Zhang and E. Bareinboim, "Equality of opportunity in classification: A causal approach," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 3675–3685.

[249] J. Komiyama, A. Takeda, J. Honda, and H. Shimao, "Nonconvex optimization for regression with fairness constraints," in *International conference on machine learning*. PMLR, 2018, pp. 2737–2746.

[250] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 2796–2806.

[251] A. Balashankar, A. Lees, C. Welty, and L. Subramanian, "What is fair? exploring pareto-efficiency for fairness constrained classifiers," *arXiv preprint arXiv:1910.14120*, 2019.

[252] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 2737–2778, 2019.

[253] A. Lamy, Z. Zhong, A. K. Menon, and N. Verma, "Noise-tolerant fair classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[254] A. Cotter, H. Jiang, and K. Sridharan, "Two-player games for efficient non-convex constrained optimization," in *Algorithmic Learning Theory*. PMLR, 2019, pp. 300–332.
[255] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu, "An algorithmic framework for fairness elicitation," *arXiv preprint arXiv:1905.10660*, 2019.
[256] A. Cotter, H. Jiang, M. R. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan, "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals." *J. Mach. Learn. Res.*, vol. 20, no. 172, pp. 1–59, 2019.
[257] M. Wick, S. Panda, and J.-B. Tristan, "Unlocking fairness: a trade-off revisited," in *NeurIPS*, 2019.
[258] A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You, "Training well-generalizing classifiers for fairness metrics and other data-dependent constraints," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1397–1405.
[259] R. Nabi, D. Malinsky, and I. Shpitser, "Learning optimal fair policies," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4674–4682.
[260] D. Xu, S. Yuan, and X. Wu, "Achieving differential privacy and fairness in logistic regression," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 594–599.
[261] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 319–328.
[262] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*. PMLR, 2019, pp. 120–129.
[263] N. Kilbertus, M. G. Rodriguez, B. Schölkopf, K. Muandet, and I. Valera, "Fair decisions despite imperfect predictions," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 277–287.
[264] M. Lohaus, M. Perrot, and U. Von Luxburg, "Too relaxed to be fair," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6360–6369.
[265] J. Ding, X. Zhang, X. Li, J. Wang, R. Yu, and M. Pan, "Differentially private and fair classification via calibrated functional mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 622–629.
[266] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Fair regression with wasserstein barycenters," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7321–7331, 2020.
[267] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. Jordan, "Robust optimization for fairness with noisy protected groups," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5190–5203, 2020.
[268] J. Cho, G. Hwang, and C. Suh, "A fair classifier using kernel density estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 088–15 099, 2020.
[269] L. Oneto, M. Donini, and M. Pontil, "General fair empirical risk minimization," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.
[270] S. Maity, D. Mukherjee, M. Yurochkin, and Y. Sun, "There is no trade-off: enforcing fairness can improve accuracy," *arXiv preprint arXiv:2011.03173*, 2020.
[271] E. Chzhen and N. Schreuder, "A minimax framework for quantifying risk-fairness trade-off in regression," *arXiv preprint arXiv:2007.14265*, 2020.
[272] M. Padala and S. Gujar, "Fnnc: Achieving fairness through neural networks," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization*, 2020.
[273] M. Scutari, F. Panero, and M. Proissl, "Achieving fairness with a simple ridge penalty," *arXiv preprint arXiv:2105.13817*, 2021.
[274] L. E. Celis, A. Mehrotra, and N. Vishnoi, "Fair classification with adversarial perturbations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8158–8171, 2021.
[275] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Fair classification with noisy protected attributes: A framework with provable guarantees," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1349–1361.
[276] A. Petrović, M. Nikolić, M. Jovanović, M. Bijanić, and B. Delibašić, "Fair classification via monte carlo policy gradient method," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104398, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197621002463
[277] K. Padh, D. Antognini, E. Lejal-Glaude, B. Faltings, and C. Musat, "Addressing fairness in classification with a model-agnostic multi-objective algorithm," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 600–609.
[278] C. Zhao, F. Chen, and B. Thuraisingham, "Fairness-aware online meta-learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2294–2304.
[279] H. Zhang, X. Chu, A. Asudeh, and S. B. Navathe, "Omnifair: A declarative system for model-agnostic group fairness in machine learning," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 2076–2088.
[280] C. Li, W. Xing, and W. Leite, "Yet another predictive model? fair predictions of students' learning outcomes in an online math learning platform," in *LAK21: 11th International Learning Analytics and Knowledge Conference*, 2021, pp. 572–578.
[281] V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, and C. Archambeau, "Fair bayesian optimization," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 854–863.
[282] A. Słowik and L. Bottou, "Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation," *arXiv preprint arXiv:2106.09467*, 2021.
[283] C. Lawless, S. Dash, O. Gunluk, and D. Wei, "Interpretable and fair boolean rule sets via column generation," *arXiv preprint arXiv:2111.08466*, 2021.
[284] Y. Choi, M. Dang, and G. Van den Broeck, "Group fairness by probabilistic modeling with latent fair decisions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 12 051–12 059.
[285] S. Park, J. Byun, and J. Lee, "Privacy-preserving fair learning of support vector machine with homomorphic encryption," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3572–3583. [Online]. Available: https://doi.org/10.1145/3485447.3512252
[286] C. Zhao, F. Mi, X. Wu, K. Jiang, L. Khan, and F. Chen, "Adaptive fairness-aware online meta-learning for changing environments," *arXiv preprint arXiv:2205.11264*, 2022.
[287] S. Boulitsakis-Logothetis, "Fairness-aware naive bayes classifier for data with multiple sensitive features," *arXiv preprint arXiv:2202.11499*, 2022.
[288] S. Hu, Z. S. Wu, and V. Smith, "Provably fair federated learning via bounded group loss," *arXiv preprint arXiv:2203.10190*, 2022.
[289] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017, pp. 5680–5689.
[290] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, "Decoupled classifiers for group-fair and efficient machine learning," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 119–133.
[291] B. Ustun, Y. Liu, and D. Parkes, "Fairness without harm: Decoupled classifiers with preference guarantees," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6373–6382.
[292] W. R. Monteiro and G. Reynoso-Meza, "Proposal of a fair voting classifier using multi-objective optimization," jul 2021.
[293] K. Kobayashi and Y. Nakao, "One-vs.-one mitigation of intersectional bias: A general method for extending fairness-aware binary classification," in *International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence*. Springer, 2021, pp. 43–54.
[294] J. Jin, Z. Zhang, Y. Zhou, and L. Wu, "Input-agnostic certified group fairness via gaussian parameter smoothing," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 340–10 361.
[295] A. Roy, V. Iosifidis, and E. Ntoutsi, "Multi-fair pareto boosting," in *International Conference on Discovery Science*. Springer, 2022.
[296] S. Liu and L. N. Vicente, "Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach," *Computational Management Science*, pp. 1–25, 2022.
[297] W. Blanzeisky and P. Cunningham, "Using pareto simulated annealing to address algorithmic bias in machine learning," *The Knowledge Engineering Review*, vol. 37, p. e5, 2022.
[298] L. Luo, W. Liu, I. Koprinska, and F. Chen, "Discrimination-aware association rule mining for unbiased data analytics," in *International Conference on Big Data Analytics and Knowledge Discovery*. Springer, 2015, pp. 108–120.

[299] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," *Advances in neural information processing systems*, vol. 29, 2016.

[300] K. D. Johnson, D. P. Foster, and R. A. Stine, "Impartial predictive modeling: Ensuring fairness in arbitrary models," *Statistical Science*, p. 1, 2016.

[301] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[302] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "Meritocratic fairness for infinite and contextual bandits," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 158–163.

[303] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1929–1938.

[304] U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum, "Multicalibration: Calibration for the (computationally-identifiable) masses," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1939–1948.

[305] S. Chiappa and W. S. Isaac, "A causal bayesian networks viewpoint on fairness," in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2018, pp. 3–20.

[306] D. Alabi, N. Immorlica, and A. Kalai, "Unleashing linear optimizers for group-fair learning and optimization," in *Conference On Learning Theory*. PMLR, 2018, pp. 2043–2066.

[307] D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: improving fairness and accuracy by learning to defer," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[308] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Model-based and actual independence for fairness-aware classification," *Data Mining and Knowledge Discovery*, vol. 32, no. 1, pp. 258–286, 2018.

[309] N. Kilbertus, A. Gascón, M. Kusner, M. Veale, K. Gummadi, and A. Weller, "Blind justice: Fairness with encrypted sensitive attributes," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2630–2639.

[310] C. Dimitrakakis, Y. Liu, D. C. Parkes, and G. Radanovic, "Bayesian fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 509–516.

[311] J. Chakraborty, T. Xia, F. M. Fahid, and T. Menzies, "Software engineering for fairness: A case study with hyperparameter optimization," *arXiv preprint arXiv:1905.05786*, 2019.

[312] A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, and A. Pentland, "Active fairness in algorithmic decision making," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 77–83.

[313] S. Chiappa, "Path-specific counterfactual fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7801–7808.

[314] V. Iosifidis and E. Ntoutsi, "Adafair: Cumulative fairness adaptive boosting," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 781–790.

[315] Y. Zhang and A. Ramesh, "Learning fairness-aware relational structures," *arXiv preprint arXiv:2002.09471*, 2020.

[316] D. Mandal, S. Deng, S. Jana, J. Wing, and D. J. Hsu, "Ensuring fairness beyond the training data," *Advances in neural information processing systems*, vol. 33, pp. 18 445–18 456, 2020.

[317] A. M. F. da Cruz, "Fairness-aware hyperparameter optimization: An application to fraud detection," 2020.

[318] V. Iosifidis and E. Ntoutsi, "Fabboo-online fairness-aware learning under class imbalance," in *International Conference on Discovery Science*. Springer, 2020, pp. 159–174.

[319] N. Martinez, M. Bertran, and G. Sapiro, "Minimax pareto fairness: A multi objective perspective," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6755–6764.

[320] A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, and J. Marques-Silva, "Towards formal fairness in machine learning," in *International Conference on Principles and Practice of Constraint Programming*. Springer, 2020, pp. 846–867.

[321] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr, "Fairfed: Enabling group fairness in federated learning," *arXiv preprint arXiv:2110.00857*, 2021.

[322] J. Wang, Y. Liu, and C. Levy, "Fair classification with group-dependent label noise," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 526–536.

[323] M. S. Ozdayi, M. Kantarcioglu, and R. Iyer, "Bifair: Training fair models with bilevel optimization," *arXiv preprint arXiv:2106.04757*, 2021.

[324] R. Islam, S. Pan, and J. R. Foulds, "Can we obtain fairness for free?" in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 586–596.

[325] S. Sharma, A. H. Gee, D. Paydarfar, and J. Ghosh, "Fair-n: Fair and robust neural networks for structured data," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 946–955. [Online]. Available: https://doi.org/10.1145/3461702.3462559

[326] J. K. Lee, Y. Bu, D. Rajan, P. Sattigeri, R. Panda, S. Das, and G. W. Wornell, "Fair selective classification via sufficiency," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6076–6086.

[327] M. Hort and F. Sarro, "Did you do your homework? raising awareness on software fairness and discrimination." ASE, 2021.

[328] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," in *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[329] A. Valdivia, J. Sánchez-Monedero, and J. Casillas, "How fair can we go in machine learning? assessing the boundaries of accuracy and fairness," *International Journal of Intelligent Systems*, vol. 36, no. 4, pp. 1619–1643, 2021.

[330] G. Wang, M. Du, N. Liu, N. Zou, and X. Hu, "Mitigating algorithmic bias with limited annotations," *arXiv preprint arXiv:2207.10018*, 2022.

[331] A. Roy and E. Ntoutsi, "Learning to teach fairness-aware deep multi-task learning," in *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2022.

[332] S. Sikdar, F. Lemmerich, and M. Strohmaier, "Getfair: Generalized fairness tuning of classification models," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 289–299.

[333] Y. Djebrouni, "Towards bias mitigation in federated learning," in *16th EuroSys Doctoral Workshop*, 2022.

[334] M. B. Short and G. O. Mohler, "A fully bayesian tracking algorithm for mitigating disparate prediction misclassification," *International Journal of Forecasting*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207022000681

[335] G. Maheshwari and M. Perrot, "Fairgrad: Fairness aware gradient descent," *arXiv preprint arXiv:2206.10923*, 2022.

[336] S. Tizpaz-Niari, A. Kumar, G. Tan, and A. Trivedi, "Fairness-aware configuration of machine learning libraries," in *Proceedings of the 44th International Conference on Software Engineering*, ser. ICSE '22. Association for Computing Machinery, 2022, p. 909–920. [Online]. Available: https://doi.org/10.1145/3510003.3510202

[337] K. Mohammadi, A. Sivaraman, and G. Farnadi, "Feta: Fairness enforced verifying, training, and predicting algorithms for neural networks," *arXiv preprint arXiv:2206.00553*, 2022.

[338] X. Gao, J. Zhai, S. Ma, C. Shen, Y. Chen, and Q. Wang, "Fairneuron: Improving deep neural network fairness with adversary games on selective neurons," in *Proceedings of the 44th International Conference on Software Engineering*, ser. ICSE '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 921–933. [Online]. Available: https://doi.org/10.1145/3510003.3510087

[339] X. Huang, Z. Li, Y. Jin, and W. Zhang, "Fair-adaboost: Extending adaboost method to achieve fair classification," *Expert Systems with Applications*, vol. 202, p. 117240, 2022.

[340] A. Candelieri, A. Ponti, and F. Archetti, "Fair and green hyperparameter optimization via multi-objective and multiple information source bayesian optimization," *arXiv preprint arXiv:2205.08835*, 2022.

[341] H. Anahideh, A. Asudeh, and S. Thirumuruganathan, "Fair active learning," *Expert Systems with Applications*, vol. 199, p. 116981, 2022.

[342] X. Li, P. Wu, and J. Su, "Accurate fairness: Improving individual fairness without trading accuracy," *arXiv preprint arXiv:2205.08704*, 2022.

[343] V. Iosifidis, A. Roy, and E. Ntoutsi, "Parity-based cumulative fairness-aware boosting," *Knowledge and Information Systems*, Jul 2022. [Online]. Available: https://doi.org/10.1007/s10115-022-01723-3

[344] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," *ACM Trans. Knowl. Discov. Data*, jul 2022, just Accepted. [Online]. Available: https://doi.org/10.1145/3551390

[345] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 99–108.

[346] W. R. Monteiro and G. Reynoso-Meza, "Proposal of a fair voting classifier using multi-objective optimization."

[347] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "The independence of fairness-aware classifiers," in *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013, pp. 849–858.

[348] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[349] M. Emmerich and A. H. Deutz, "A tutorial on multiobjective optimization: fundamentals and evolutionary methods," *Natural computing*, vol. 17, no. 3, pp. 585–609, 2018.

[350] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowledge and Information Systems*, vol. 54, no. 1, pp. 95–122, 2018.

[351] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[352] G. Morina, V. Oliinyk, J. Waton, I. Marusic, and K. Georgatzis, "Auditing and achieving intersectional fairness in classification problems," *arXiv preprint arXiv:1911.01468*, 2019.

[353] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 247–254.

[354] K. Kanamori and H. Arimura, "Fairness-aware edit of thresholds in a learned decision tree using a mixed integer programming formulation," in *The 33rd Annual Conference of the Japanese Society for Artificial Intelligence (2019)*. The Japanese Society for Artificial Intelligence, 2019, pp. 3Rin211–3Rin211.

[355] Y. Savani, C. White, and N. S. Govindarajulu, "Intra-processing methods for debiasing neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2798–2810, 2020.

[356] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Fair regression via plug-in estimator and recalibration with statistical guarantees," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 137–19 148, 2020.

[357] P. Awasthi, M. Kleindessner, and J. Morgenstern, "Equalized odds postprocessing under imperfect group information," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1770–1780.

[358] N. Schreuder and E. Chzhen, "Classification with abstention but without disparities," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 1227–1236.

[359] K. Kanamori and H. Arimura, "Fairness-aware decision tree editing based on mixed-integer linear optimization," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 36, no. 4, pp. B–L13_1, 2021.

[360] A. Mishler, E. H. Kennedy, and A. Chouldechova, "Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 386–400. [Online]. Available: https://doi.org/10.1145/3442188.3445902

[361] M. Du, S. Mukherjee, G. Wang, R. Tang, A. Awadallah, and X. Hu, "Fairness via representation neutralization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 091–12 103, 2021.

[362] P. A. Grabowicz, N. Perello, and A. Mishra, "Marrying fairness and explainability in supervised learning," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1905–1916.

[363] J. Zhang, I. Beschastnikh, S. Mechtaev, and A. Roychoudhury, "Fair decision making via automated repair of decision trees," in *International Workshop on Equitable Data and Technology (FairWare '22)*, 2022.

[364] N. Mehrabi, U. Gupta, F. Morstatter, G. V. Steeg, and A. Galstyan, "Attributing fair decisions with attention interventions," in *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Seattle, U.S.A.: Association for Computational Linguistics, Jul. 2022, pp. 12–25. [Online]. Available: https://aclanthology.org/2022.trustnlp-1.2

[365] Z. Wu and J. He, "Fairness-aware model-agnostic positive and unlabeled learning," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1698–1708. [Online]. Available: https://doi.org/10.1145/3531146.3533225

[366] R. Marcinkevics, E. Ozkan, and J. E. Vogt, "Debiasing deep chest x-ray classifiers using intra- and post-processing methods," in *Machine Learning for Healthcare Conference*. PMLR, 2022.

[367] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring discrimination in socially-sensitive decision records," in *Proceedings of the 2009 SIAM international conference on data mining*. SIAM, 2009, pp. 581–592.

[368] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 144–152.

[369] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 107–118.

[370] J. Liu, J. Li, L. Liu, T. D. Le, F. Ye, and G. Li, "Fairmod-making predictive models discrimination aware," *arXiv preprint arXiv:1811.01480*, 2018.

[371] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Leveraging labeled and unlabeled data for consistent fair binary classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[372] P. K. Lohia, K. Natesan Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2847–2851.

[373] I. Alabdulmohsin, "Fair classification via unconstrained optimization," *arXiv preprint arXiv:2005.14621*, 2020.

[374] I. M. Alabdulmohsin and M. Lucic, "A near-optimal algorithm for debiasing trained machine learning models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8072–8084, 2021.

[375] D. Nguyen, S. Gupta, S. Rana, A. Shilton, and S. Venkatesh, "Fairness improvement for black-box classifiers with gaussian process," *Information Sciences*, vol. 576, pp. 542–556, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025521006861

[376] P. Lohia, "Priority-based post-processing bias mitigation for individual and group fairness," *arXiv preprint arXiv:2102.00417*, 2021.

[377] T. Jang, P. Shi, and X. Wang, "Group-aware threshold adaptation for fair classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6988–6995.

[378] P. Snel and S. van Otterloo, "Practical bias correction in neural networks: a credit default prediction case study," *Computers and Society Research Journal*, no. 3, 2022.

[379] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. W. Michalak, S. Asoodeh, and F. P. Calmon, "Beyond adult and compas: Fairness in multi-class prediction," *arXiv preprint arXiv:2206.07801*, 2022.

[380] X. Zeng, E. Dobriban, and G. Cheng, "Fair bayes-optimal classifiers under predictive parity," *arXiv preprint arXiv:2205.07182*, 2022.

[381] ——, "Bayes-optimal classifiers under group fairness," *arXiv preprint arXiv:2202.09724*, 2022.

[382] X. Yin and J. Han, "Cpar: Classification based on predictive association rules," in *Proceedings of the 2003 SIAM international conference on data mining*. SIAM, 2003, pp. 331–335.

[383] B. Ghai, M. Mishra, and K. Mueller, "Cascaded debiasing: Studying the cumulative effect of multiple fairness-enhancing interventions," *arXiv preprint arXiv:2202.03734*, 2022.

[384] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6478–6490, 2021.

[385] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[386] M. Redmond and A. Baveja, "A data-driven software tool for enabling cooperative information sharing among police departments," *European Journal of Operational Research*, vol. 141, no. 3, pp. 660–678, 2002.

[387] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[388] L. F. Wightman, "Lsac national longitudinal bar passage study. lsac research report series." 1998.

[389] I.-C. Yeh and C.-h. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of

credit card clients," *Expert systems with applications*, vol. 36, no. 2, pp. 2473–2480, 2009.
[390] "Dutch central bureau for statistics volkstelling," http://easy.dans.knaw.nl/dms, 2001, retrieved on June 12, 2022.
[391] "The heritage health prize dataset," https://www.kaggle.com/c/hhp, 2017, retrieved on June 12, 2022.
[392] "Medical expenditure panel survey dataset," https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192, 2016, retrieved on June 12, 2022.
[393] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban, "The five factor model of personality and evaluation of drug consumption risk," in *Data science*. Springer, 2017, pp. 231–242.
[394] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.
[395] "National longitudinal surveys of youth data set," www.bls.gov/nls/, 2019, retrieved on June 12, 2022.
[396] "Stop, question and frisk dataset," http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page, 2017, retrieved on June 12, 2022.
[397] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision support systems*, vol. 47, no. 4, pp. 547–553, 2009.
[398] Supreme Court of the United States, *Ricci v. DeStefanoo*, 2009, vol. 557.
[399] "Home credit default risk," https://www.kaggle.com/c/home-credit-default-risk, 2018, retrieved on June 12, 2022.
[400] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
[401] T. Kehrenberg, Z. Chen, and N. Quadrianto, "Tuning fairness by marginalizing latent target labels," *stat*, vol. 1050, p. 10, 2019.
[402] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.
[403] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
[404] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–16.
[405] C. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Müller, and P. Kessel, "Fairwashing explanations with off-manifold detergent," in *International Conference on Machine Learning*. PMLR, 2020, pp. 314–323.
[406] J. Sylvester and E. Raff, "Trimming the thorns of ai fairness research." *IEEE Data Eng. Bull.*, vol. 43, no. 4, pp. 74–84, 2020.
[407] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.