Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey

MAX HORT, Simula Research Laboratory, Norway
ZHENPENG CHEN, University College London, United Kingdom
JIE M. ZHANG, King's College London, United Kingdom
MARK HARMAN, University College London, United Kingdom
FEDERICA SARRO, University College London, United Kingdom

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 technique 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.

CCS Concepts: • **Computing methodologies → Supervised learning by classification**; *Artificial intelligence.*

Additional Key Words and Phrases: fairness, bias mitigation, debiasing, fairness-aware machine learning, classification

ACM Reference Format:
Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2018. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. In . ACM, New York, NY, USA, 52 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Machine Learning (ML) has been increasingly popular in recent years, both in the diversity and importance of applications [75]. ML is used in a variety of critical applications such as justice risk assessments [24, 37], job recommendations [411], and autonomous driving [226].

While ML systems have the advantage to relieve humans from tedious tasks and are able to perform complex calculations at a higher speed [286], they are only as good as the data on which they are trained [33]. ML algorithms, which are never designed to intentionally incorporate bias, run the risk of replicating or even amplifying bias present in real-world data [33, 282, 346]. 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 [186, 187, 255]. For

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

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 and there are many examples of this in the literature. For instance, bias has been found in advertisement and recruitment processes [92, 411], affecting university admissions [40] and human rights [255]. Not only is such a biased behaviour undesired, but it can fall under regulatory control and risk the violation of anti-discrimination laws [66, 282, 310], 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 [211].

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 [24].

To reduce the degree of bias that such systems exhibit, practitioners use three types of bias mitigation methods [122]:

- **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.

In this survey, we use the terms "bias mitigation" and "fairness improvement" interchangeably and treat fairness as the absence of bias.

There has been a growing interest in fairness research, including definitions, measurements, and improvements of ML models [69, 70, 75, 107, 286]. 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) [348].

Despite the large amount of existing bias mitigation methods and surveys on fairness research, as Pessach and Shmueli [286] 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?" [286]

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 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 most comprehensive survey to systematically search and cover bias mitigation methods and their empirical evaluation. To summarize, the contributions 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: https://docs.google.com/spreadsheets/d/1kOmbKLMiFgHRSXvgM-O8OW4YKeDIGN0cPPeCGQOMnnA/edit?usp=sharing.

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. Opportunities and challenges that the field of fairness research and bias mitigation methods face are discussed in Section 8. Section 9 provides recommendations to practitioners, distilled from the collected publications. Section 10 concludes this survey.

2 RELATED SURVEYS

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

Mehrabi et al. [255] and Pessach and Shmueli [286] provided an overview of bias and discrimination types, fairness definitions and metrics, bias mitigation methods, and existing datasets. For example, Pessach and Shmueli [287] listed the datasets and metrics used by 27 bias mitigation methods. A similar focus has been pursued by Dunkelau and Leuschel [107], 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 [51], who also presented fairness metrics and fairness platforms.

A detailed collection of prominent fairness definitions for classification problems is provided by Verma and Rubin [348]. Similarly, Žliobaite [415] surveyed measures for indirect discrimination for ML. While these collections describe current metrics used to determine the fairness of ML models, Hutchinson and Mitchell [155] drew parallels from fairness research in the 1960s and 1970s concerning test fairness, for education and hiring, to current advances. Similar to modern metrics and evaluation approaches, past work considered fairness with regards to individuals and groups, or the use of confusion matrix measures (Section 6).

In addition to the surveys on fairness metrics, Le Quy et al. [218] 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 [405], criminal justice [37], graph representations [404], ML testing [396], Software Engineering [67, 334], or Natural Language Processing [43, 336].

While previous surveys focused on ML classification, and some mentioned 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 ours are provided by Dunkelau and Leuschel [107], and Pessach and Shmueli [287].

Dunkelau and Leuschel [107] 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 [287] listed the datasets and metrics used by 27 bias mitigation methods, they did 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 [107], and providing information on datasets and metrics similar to Pessach and Shmueli [286]), 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. [255] : 24 bias mitigation methods;
- Pessach and Shmueli [287]: 30 bias mitigation methods;
- Dunkelau and Leuschel [107]: 40 bias mitigation methods;
- Caton and Haas [51]: 70 bias mitigation methods.

In total, we collect 100 unique publications with 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;
- 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* keywords ensure 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).

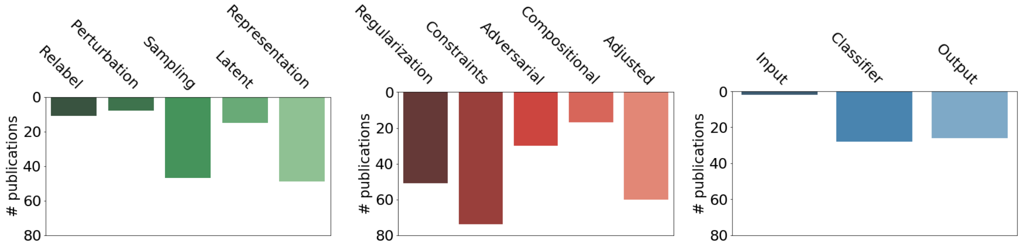
To ensure that irrelevant publications are excluded from the search results, we manually check publications in three filtration stages [250]:

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 [363]. 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.

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



(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.

3.2 Selected Publications

In total, we gathered 341 publications over the different stages of our search procedure. Table 2 summarises the results of the 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 entries.

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 32 publications. The amount of publications found for each step of the search is listed in Table 1.

The publication distribution per year and venue type is illustrated in Figure 2. We categorized the 341 publications in five venue types, in line with the categories by Soremekun et al. [334]: Artificial Intelligence (AI), Data, Fairness, Software Engineering (SE), other. Note that the category "other" consists of 100 publications, 68 of which are published on arXiv. The category SE combines publications form Software Engineering, Programming Language and Security venues. From this figure, we can see that there is an increasing interest in bias mitigation methods and a steady increase of publications over the years. In particular, we observe a huge jump in the number of

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.

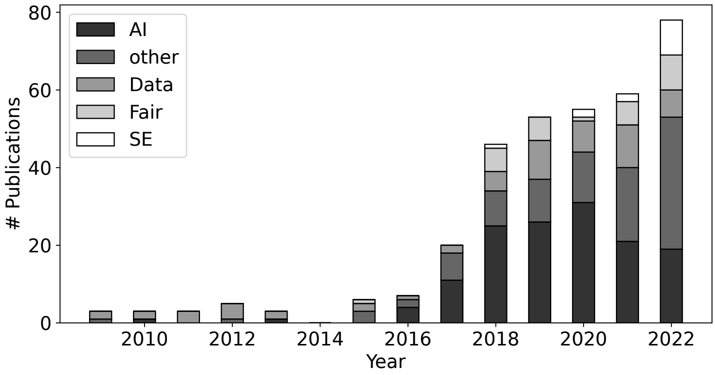


Fig. 2. Number of publications per year venue type.

publications in 2018, more than doubling the number of publications from 2017 (i.e., from 20 to 46 publications). Prior years, from 2009-2016, have seen less than 10 publications each.

The publication venues with the highest number of publications are: NeurIPS (38 publications), ICML (27 publications), AAAI (18 publications), FAccT (13 publications), AIES (12 publications).

3.3 Visibility

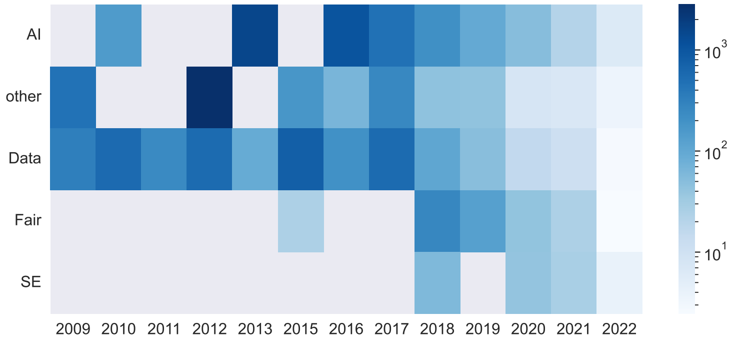


Fig. 3. Average number of citations per year and venue type.

In this section, we address the visibility of bias mitigation methods within the context of research venues and publications. We use the amount of publications and number of citations as a proxy for the visibility of bias mitigation methods across different publication venues.[1]

As shown in Figure 2, there is an increasing trend in the number of publications on bias mitigation methods per year, which supports the claim that the visibility and relevance of bias mitigation is growing. Among the five venue types (AI, Data, Fairness, SE, other), bias mitigation methods exhibit the highest visibility in terms of number of publications for AI (139 publications), data (59 publications) and other venues (most notably arXiv with 68 publications). The past five years, from 2018 onwards, saw an uptake of bias mitigation methods in a wider range of venues, with the inclusion of bias mitigation methods in Software Engineering venues and the creation of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT),[2] as well as specialised venues co-located with well-renowned international conference such as the IEEE/ACM International Workshop on Equitable Data & Technology (FairWare) at the International Conference of Software Engineering.[3]

Figure 3 provides a closer look at the average number of citations of publications for each venue type. We can see that publications from early years of bias mitigation methods have a high average visibility (i.e., number of citations). A reason for this can be found in the low number of publications, with only 3-7 publications yearly from 2009-2016, and the relevance of such publications to be the foundation of proceeding work. Data venues published bias mitigation methods consistently, every year from 2009 to 2022. While Fairness and SE venues have fewer publications per year, the respective papers achieve a high visibility, frequently with a higher average number of citations than Data and AI venues for the same years. Thereby, the highest average number of citations was achieved by publications in fairness venues in 2018, 2019 and 2021.

Among the most cited publications (19 of which publications have been cited more than 500 times) only two have not been published in AI or data venues. This includes the work by Dwork et al. [108] (published in the proceedings of the 3rd innovations in theoretical computer science conference) and Zhang et al. [391] (published at the AAAI/ACM Conference on AI, Ethics, and Society). We note that 10 out of the 15 most cited works have publicly available implementations in fairness frameworks [31, 35, 41].

[1] We obtained the number of citations for each publication from Google Scholar on February 24th, 2023, and included them in our online repository [25].
[2] https://facctconference.org/index.html
[3] https://dblp.org/db/conf/fairware-ws/index.html

3.4 Limitations

This survey focuses on investigating the fairness of ML models from an algorithmic point of view. While fairness is a multi-disciplinary field of research, and has been addressed by various communities, including law [44], health studies [271], and criminal justice [32], we focus on fairness and bias as exhibited by ML models.

Moreover, our search procedure is designed to find publications that mitigate bias for tabular datasets. This does not mean that we exclude a priori relevant publications if they have been published at Computer Vision or Natural Language Processing venues. In fact, such publications are considered in our survey if bias mitigation for tabular datasets is addressed in a part of the publication, whereas bias mitigation methods that are solely applied for visual or textual tasks are not included.

Furthermore, we note that the overview presented herein is based on bias mitigation methods as proposed by the research community, often applied to publicly available datasets. While these datasets can be based on real-world scenarios, results might not transfer to real-world applications [32, 98].

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 [122]. Moreover, we organize methods in categories (i.e., the bias mitigation approach). For this, we follow taxonomies devised by Dunkelau and Leuschel [107], as well as Caton and Haas [51]. Figure 1 illustrates the 13 categories we use.

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). Note that a single publication can apply up to three different types of bias mitigation methods and can be counted multiple times, 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.

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 lists 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 datasets is "massaging", proposed by Kamiran and Calders [182]. 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. Typically, an equal amount of instances with positive and negative labels are selected, according to their rank and their labels are switched.

Massaging has later been extended by Kamiran and Calders [184], and Calders et al. [46]. Moreover, Žliobaite et al. [416] created a related method called "local massaging". "Massaging" has also been applied by other work [163, 398].

Another relabelling approach was proposed by Luong et al. [241], who relabelled instances based on their k-nearest neighbours, such that similar individuals receive similar labels.

Table 3. Publications on Pre-processing bias mitigation methods.

Type	Authors [Ref]	Year	Venue
Relabel	Calders et al. [46]	2009	ICDMW
	Kamiran and Calders [182]	2009	ICCCC
	Žliobaite et al. [416]	2011	ICDM
	Luong et al. [241]	2011	KDD
	Hajian and Domingo-Ferrer [142]	2012	TKDE
	Kamiran and Calders [184]	2012	KAIS
	Zhang et al. [398]	2018	IJCAI
	Iosifidis et al. [163]	2019	DEXA
	Seker et al. [324]	2022	HTI
	Sun et al. [335]	2022	EuroS&P
	Alabdulmohsin et al. [16]	2022	arXiv
Perturbation	Hajian and Domingo-Ferrer [142]	2012	TKDE
	Feldman et al. [117]	2015	KDD
	Lum and Johndrow [239]	2016	arXiv
	Wang et al. [353]	2018	NeurIPS
	Johndrow and Lum [172]	2019	Ann Appl Stat
	Wang et al. [352]	2019	ICML
	Li et al. [225]	2022	SSRN
	Li et al. [228]	2022	ICSE
Sampling	Calders et al. Calders et al. [46]	2009	ICDMW
	Kamiran and Calders [183]	2010	BNAIC
	Žliobaite et al. [416]	2011	ICDM
	Kamiran and Calders [184]	2012	KAIS
	Zhang et al. [397]	2017	IJCAI
	Chen et al. [65]	2018	NeurIPS
	Iosifidis and Ntoutsi [159]	2018	report
	Xu et al. [372]	2018	Big Data
	Krasanakis et al. [210]	2018	TheWebConf
	Abusitta et al. [8]	2019	arXiv
	Xu et al. [370]	2019	IJCAI
	Zelaya et al. [387]	2019	KDD
	Salimi et al. [318]	2019	MOD
	Iosifidis et al. [163]	2019	DEXA
	Iosifidis et al. [158]	2019	Big Data
	Xu et al. [373]	2019	Big Data
	Abay et al. [7]	2020	arXiv
	Hu et al. [152]	2020	DS
	Chakraborty et al. [61]	2020	FSE
	Jiang and Nachum [168]	2020	AISTATS
	Sharma et al. [326]	2020	AIES
	Celis et al. [55]	2020	ICML
	Morano [262]	2020	Thesis
	Yan et al. [374]	2020	CIKM
	Chuang and Mroueh [76]	2021	ICLR
	Salazar et al. [317]	2021	IEEE Access
	Zhang et al. [399]	2021	PAKDD
	Yu [379]	2021	arXiv
	Iofinova et al. [157]	2021	arXiv
	Roh et al. [308]	2021	NeurIPS
	Du and Wu [104]	2021	CIKM
	Singh et al. [330]	2021	MAKE
	Amend and Spurlock [21]	2021	JCSC
	Jang et al. [167]	2021	AAAI
	Verma et al. [347]	2021	arXiv
	Chakraborty et al. [59]	2021	FSE
	Cruz et al. [88]	2021	ICDM
	Wang et al. [356]	2022	ICML
	Pentyala et al. [283]	2022	arXiv
	Rajabi and Garibay [298]	2022	MAKE
	Sun et al. [335]	2022	EuroS&P
	Dablain et al. [90]	2022	arXiv
	Chen et al. [68]	2022	FSE
	Li and Liu [224]	2022	PMLR
	Chakraborty et al. [60]	2022	FairWARE
	Almuzaini et al. [20]	2022	FAccT
	Chai and Wang [58]	2022	ICML
Latent	Calders and Verwer [48]	2010	DMKD
	Kilbertus et al. [201]	2017	NeurIPS
	Gupta et al. [139]	2018	arXiv
	Madras et al. [244]	2019	FAccT
	Oneto et al. [275]	2019	AIES
	Wei et al. [360]	2020	PMLR
	Kehrenberg et al. [197]	2020	Front. Artif. Intell.
	Grari et al. [136]	2021	arXiv
	Chen et al. [64]	2022	arXiv
	Liang et al. [229]	2022	arXiv
	Jung et al. [178]	2022	CVPR
	Diana et al. [96]	2022	FAccT
	Chakraborty et al. [60]	2022	FairWARE
	Wu et al. [366]	2022	CLeaR
	Suriyakumar et al. [338]	2022	arXiv
Representation	Zemel et al. [388]	2013	ICML
	Edwards and Storkey [110]	2015	arXiv
	Louizos et al. [237]	2016	ICLR
	Pérez-Suay et al. [284]	2017	ECML PKDD
	Calmon et al. [49]	2017	NeurIPS
	Hacker and Wiedemann [141]	2017	arXiv
	Komiyama and Shimao [208]	2017	arXiv
	Xie et al. [369]	2017	NeurIPS
	McNamara et al. [253]	2017	arXiv
	du Pin Calmon et al. [105]	2018	IEEE J Sel
	Grgić-Hlača et al. [138]	2018	AAAI
	Madras et al. [243]	2018	ICML
	Samadi et al. [319]	2018	NeurIPS
	Quadrianto et al. [294]	2018	arXiv
	Moyer et al. [265]	2018	NeurIPS
	Song et al. [333]	2019	AISTATS
	Gordaliza et al. [132]	2019	ICML
	Quadrianto et al. [295]	2019	CVPR
	Creager et al. [87]	2019	ICML
	Wang and Huang [358]	2019	arXiv
	Lahoti et al. [214]	2019	ICDE
	Feng et al. [118]	2019	arXiv
	Lahoti et al. [215]	2019	VLDB
	Zhao et al. [409]	2020	ICLR
	Tan et al. [340]	2020	AISTATS
	Jaiswal et al. [165]	2020	AAAI
	Zehlike et al. [386]	2020	DMKD
	Sarhan et al. [320]	2020	ECCV
	Madhavan and Wadhwa [242]	2020	CIKM
	Kim and Cho [203]	2020	AAAI
	Ruoss et al. [313]	2020	NeurIPS
	Fong et al. [121]	2021	arXiv
	Gupta et al. [140]	2021	AAAI
	Zhu et al. [414]	2021	ICCV
	Grari et al. [134]	2021	ECML PKDD
	Salazar et al. [316]	2021	VLDB
	Oh et al. [272]	2022	arXiv
	Agarwal and Deshpande [13]	2022	FAccT
	Wu et al. [365]	2022	arXiv
	Shui et al. [328]	2022	arXiv
	Qi et al. [291]	2022	arXiv
	Balunović et al. [30]	2022	ICLR
	Kairouz et al. [179]	2022	T-IFS
	Liu et al. [231]	2022	Neural Process. Lett.
	Cerrato et al. [57]	2022	arXiv
	Kamani et al. [181]	2022	Mach. Learn.
	Rateike et al. [300]	2022	FAccT
	Galhotra et al. [125]	2022	SIGMOD
	Kim and Cho [204]	2022	Neurocomputing

Feldman et al. [117] 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). Johndrow and Lum [172], Lum and Johndrow [239] used conditional models for perturbation, which allowed for modification of multiple variables (continuous or discrete). Li et al. [225] 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 [117, 172, 239], Wang et al. [352, 353] considered only the unprivileged group for perturbation, seeking to resolve disparity by improving the performance of the unprivileged group. Hajian and Domingo-Ferrer [142] 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 adjusted by reweighing their importance [7, 20, 46, 55, 58, 104, 163, 184, 224, 283, 379].

Reweighing was first introduced by Calders et al. [46]. 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 [184].

Jiang and Nachum [168] and Krasanakis et al. [210] 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 and Liu [224] assigned individual weights to instances of the training data.

Other sampling strategies include the removal of data points (downsampling) [61, 68, 88, 157, 308, 318, 347, 356, 399] or the addition of new data points (upsampling). Popular methods for upsamplig are oversampling for duplicating instances of the minority group [21, 159, 262, 387] and the use of SMOTE [63]. SMOTE does not duplicate instances but generates synthetic ones in the neighborhood of the minority group [59, 60, 90, 159, 262, 317, 330, 374, 387].

To sample datapoints, uniform [184] and preferential [152, 183, 184, 387, 416] strategies have been followed, where preferential sampling changes the distribution of instances close to the decision boundary.

Xu et al. [370, 372, 373] used a generative approach to generate discrimination-free data for training [8, 167, 298]. Zhang et al. [397] used causal networks to create a new dataset. The initial dataset is used to create a causal network, which is then modified to reduce discrimination. The debiased causal network is used to generate a new dataset. Sharma et al. [326] 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 variables describe the augmentation of training data with additional features that are preferably unbiased. In previous work, latent variables have been used to represent labels [197, 360] and group memberships (i.e., protected or unprotected group) [60, 64, 96, 136, 139, 178, 229, 275, 338], and are frequently considered when dealing with causal graphs [136, 201, 244].

For instance, Calders and Verwer [48] 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. [139] 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.

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

The first representation learning approach for bias mitigation was Learning Fair Representations (LFR), proposed by Zemel et al. [388]. 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 [49, 105, 132, 141, 214, 215, 253, 265, 328, 333, 386]. Other used techniques are:

- adversarial learning [110, 118, 134, 165, 179, 203, 243, 291, 313, 369, 409, 414];
- variational autoencoders [87, 231, 237, 272, 300];
- adversarial variational autoencoder [365];
- normalizing flows [30, 57];
- dimensionality reduction [181, 284, 319, 340];
- residuals [208];
- contrastive learning [140];
- neural style transfer [294, 295].

Another method for improving the fairness of data representations is the removal [138, 242, 358] or addition of features [121, 125, 316]. Grgić-Hlača et al. [138] investigated fairness while using different sets of features, thereby making training feature choices. Madhavan and Wadhwa [242] removed discriminating features from the training data. Salazar et al. [316] applied feature creation techniques which apply nonlinear transformation and 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 4, Table 5 for more details) that apply in-processing methods. For more details on in-processing methods, we refer to the survey by Wan et al. [350], 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 a 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. [52] 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 [185, 299, 354, 400–403]. Traditionally, leaves are iteratively split to achieve an improvement in accuracy. To improve fairness while training, Kamiran et al. [185] considered fairness in addition to accuracy when leaf splitting. They applied three splitting strategies:

(1) only allow non-discriminatory splits;

Table 4. Publications on In-processing bias mitigation methods.

Type	Authors [Ref]	Year	Venue
Regularization	Kamiran et al. [185]	2010	ICDM
	Kamishima et al. [191]	2011	ICDMW
	Kamishima et al. [188]	2012	ECML PKDD
	Ristanoski et al. [305]	2013	CIKM
	Fish et al. [119]	2015	FATML
	Pérez-Suay et al. [284]	2017	ECML PKDD
	Bechavod and Ligett [34]	2017	arXiv
	Berk et al. [36]	2017	arXiv
	Quadrianto and Sharmanska [293]	2017	NeurIPS
	Raff et al. [297]	2018	AIES
	Enni and Assent [112]	2018	ICDM
	Goel et al. [130]	2018	AAAI
	Zhang et al. [401]	2019	ICDMW
	Mary et al. [252]	2019	ICML
	Beutel et al. [38]	2019	AIES
	Huang and Vishnoi [153]	2019	ICML
	Aghaei et al. [14]	2019	AAAI
	Zhang and Ntoutsi [400]	2019	IJCAI
	Keya et al. [198]	2020	arXiv
	Kim et al. [202]	2020	ICML
	Jiang et al. [169]	2020	UAI
	Di Stefano et al. [95]	2020	arXiv
	Abay et al. [7]	2020	arXiv
	Baharlouei et al. [28]	2020	ICLR
	Liu et al. [233]	2020	Preprint
	Kamani [180]	2020	Thesis
	Ravichandran et al. [301]	2020	arXiv
	Tavakol [342]	2020	SIGIR
	Romano et al. [309]	2020	NeurIPS
	Hickey et al. [147]	2020	ECML PKDD
	Wang et al. [359]	2021	SIGKDD
	Chuang and Mroueh [76]	2021	ICLR
	Lowy et al. [238]	2021	arXiv
	Zhang and Weiss [402]	2021	ICDM
	Grari et al. [135]	2021	IJCAI
	Yurochkin and Sun [381]	2021	ICLR
	Zhao et al. [412]	2021	arXiv
	Ranzato et al. [299]	2021	CIKM
	Mishler and Kennedy [257]	2021	arXiv
	Kang et al. [194]	2021	arXiv
	Sun et al. [335]	2022	EuroS&P
	Zhao et al. [413]	2022	WSDM
	Wang et al. [354]	2022	CAV
	Deng et al. [94]	2022	arXiv
	Lee et al. [220]	2022	Entropy
	Zhang and Weiss [403]	2022	AAAI
	Jiang et al. [170]	2022	ICLR
	Lee et al. [219]	2022	ICASSP
	Do et al. [101]	2022	ICML
	Patil and Purcell [280]	2022	Future Internet
	Kim and Cho [204]	2022	Neurocomputing
Constraints	Dwork et al. [108]	2012	ITCS
	Calders et al. [47]	2013	ICDM
	Fukuchi and Sakuma [124]	2015	arXiv
	Fukuchi et al. [123]	2015	IEICE Trans. Inf.& Syst.
	Goh et al. [131]	2016	NeurIPS
	Woodworth et al. [364]	2017	COLT
	Zafar et al. [382]	2017	TheWebConf
	Corbett-Davies et al. [81]	2017	KDD
	Zafar et al. [385]	2017	AISTATS
	Komiyama and Shimao [208]	2017	arXiv
	Zafar et al. [384]	2017	NeurIPS
	Quadrianto and Sharmanska [293]	2017	NeurIPS
	Russell et al. [314]	2017	NeurIPS
	Kilbertus et al. [201]	2017	NeurIPS
	Agarwal et al. [11]	2018	ICML
	Kim et al. [205]	2018	NeurIPS
	Narasimhan [268]	2018	AISTATS
	Gillen et al. [129]	2018	NeurIPS
	Grgić-Hlača et al. [138]	2018	AAAI
	Heidari et al. [146]	2018	NeurIPS
	Kearns et al. [195]	2018	ICML
	Zhang and Bareinboim [394]	2018	AAAI
	Gupta et al. [139]	2018	arXiv
	Olfat and Aswani [273]	2018	AISTATS
	Zhang and Bareinboim [393]	2018	NeurIPS
	Komiyama et al. [209]	2018	ICML
	Wu et al. [367]	2018	arXiv
	Donini et al. [102]	2018	NeurIPS
	Farnadi et al. [115]	2018	AIES
	Nabi and Shpitser [267]	2018	AAAI
	Goel et al. [130]	2018	AAAI
	Wick et al. [361]	2019	NeurIPS
	Celis et al. [52]	2019	FAccT
	Cotter et al. [84]	2019	ICML
	Balashankar et al. [29]	2019	arXiv
	Agarwal et al. [12]	2019	ICML
	Nabi et al. [266]	2019	ICML
	Cotter et al. [86]	2019	ALT
	Oneto et al. [275]	2019	AIES
	Cotter et al. [85]	2019	JMLR
	Jung et al. [177]	2019	arXiv
	Lamy et al. [216]	2019	NeurIPS
	Xu et al. [371]	2019	TheWebConf
	Zafar et al. [383]	2019	JMLR
	Wang et al. [357]	2020	NeurIPS
	Chzhen and Schreuder [80]	2020	arxiv
	Lohaus et al. [234]	2020	ICML
	Kilbertus et al. [200]	2020	AISTATS
	Ding et al. [99]	2020	AAAI
	Maity et al. [247]	2020	arXiv
	Cho et al. [73]	2020	NeurIPS
	Padala and Gujar [277]	2020	IJCAI
	Oneto et al. [274]	2020	IJCNN
	Chzhen et al. [79]	2020	NeurIPS
	Celis et al. [53]	2021	PMLR
	Celis et al. [56]	2021	NeurIPS
	Słowik and Bottou [331]	2021	arXiv
	Li et al. [223]	2021	LAK
	Scutari et al. [323]	2021	arXiv
	Padh et al. [278]	2021	UAI
	Zhang et al. [392]	2021	MOD
	Zhao et al. [407]	2021	KDD
	Petrović et al. [289]	2021	Eng. Appl. Artif. Intell.
	Perrone et al. [285]	2021	AIES
	Choi et al. [74]	2021	AAAI
	Du and Wu [104]	2021	CIKM
	Lawless et al. [217]	2021	arXiv
	Mishler and Kennedy [257]	2021	arXiv
	Park et al. [279]	2022	WWW
	Wang et al. [354]	2022	CAV
	Zhao et al. [408]	2022	KDD
	Boulitsakis-Logothetis [45]	2022	arXiv
	Hu et al. [151]	2022	arXiv
	Wu et al. [366]	2022	CLeaR

Table 5. Publications on In-processing bias mitigation methods - Part 2.

Type	Authors [Ref]	Year	Venue
Adversarial	Beutel et al. [39]	2017	arXiv
	Agarwal et al. [11]	2018	ICML
	Gillen et al. [129]	2018	NeurIPS
	Raff and Sylvester [296]	2018	DSAA
	Wadsworth et al. [349]	2018	arXiv
	Kearns et al. [195]	2018	ICML
	Zhang et al. [391]	2018	AIES
	Adel et al. [9]	2019	AAAI
	Beutel et al. [38]	2019	AIES
	Sadeghi et al. [315]	2019	ICCV
	Zhao and Gordon [410]	2019	NeurIPS
	Xu et al. [373]	2019	Big Data
	Grari et al. [137]	2019	ICDM
	Celis and Keswani [54]	2019	arXiv
	Garcia de Alford et al. [127]	2020	SMU DSR
	Yurochkin et al. [380]	2020	ICLR
	Roh et al. [306]	2020	ICML
	Delobelle et al. [93]	2020	ASE
	Rezaei et al. [303]	2020	AAAI
	Lahoti et al. [213]	2020	NeurIPS
	Grari et al. [136]	2021	arXiv
	Grari et al. [135]	2021	IJCAI
	Amend and Spurlock [21]	2021	JCSC
	Rezaei et al. [304]	2021	AAAI
	Chen et al. [64]	2022	arXiv
	Liang et al. [229]	2022	arXiv
	Tao et al. [341]	2022	FSE
	Petrović et al. [288]	2022	Neurocomputing
	Yang et al. [375]	2022	medRxiv
	Yazdani-Jahromi et al. [376]	2022	arXiv
Compositional	Calders and Verwer [48]	2010	DMKD
	Pleiss et al. [290]	2017	NeurIPS
	Dwork et al. [109]	2018	FAccT
	Ustun et al. [344]	2019	ICML
	Oneto et al. [275]	2019	AIES
	Iosifidis et al. [158]	2019	Big Data
	Monteiro and Reynoso-Meza [261]	2021	PLM
	Ranzato et al. [299]	2021	CIKM
	Mishler and Kennedy [257]	2021	arXiv
	Kobayashi and Nakao [207]	2021	DiTTEt
	Jin et al. [171]	2022	ICML
	Chen et al. [68]	2022	FSE
	Roy et al. [311]	2022	DS
	Liu and Vicente [232]	2022	CMS
	Blanzeisky and Cunningham [42]	2022	Knowl Eng Rev
	Boulitsakis-Logothetis [45]	2022	arXiv
	Suriyakumar et al. [338]	2022	arXiv

Type	Authors [Ref]	Year	Venue
Adjusted	Luo et al. [240]	2015	DaWaK
	Joseph et al. [176]	2016	NeurIPS
	Johnson et al. [174]	2016	Stat Sci
	Kusner et al. [212]	2017	NeurIPS
	Joseph et al. [175]	2018	AIES
	Hashimoto et al. [144]	2018	ICML
	Madras et al. [245]	2018	NeurIPS
	Alabi et al. [18]	2018	COLT
	Hébert-Johnson et al. [145]	2018	ICML
	Chiappa and Isaac [72]	2018	IFIP
	Kilbertus et al. [199]	2018	ICML
	Kamishima et al. [190]	2018	DMKD
	Dimitrakakis et al. [97]	2019	AAAI
	Chiappa [71]	2019	AAAI
	Noriega-Campero et al. [270]	2019	AIES
	Chakraborty et al. [62]	2019	arXiv
	Madras et al. [244]	2019	FAccT
	Iosifidis and Ntoutsi [160]	2019	CIKM
	Mandal et al. [248]	2020	NeurIPS
	Kilbertus et al. [200]	2020	AISTATS
	Martinez et al. [251]	2020	ICML
	Iosifidis and Ntoutsi [161]	2020	DS
	Liu et al. [233]	2020	Preprint
	Hu et al. [152]	2020	DS
	da Cruz [89]	2020	Thesis
	Chakraborty et al. [61]	2020	FSE
	Kamani [180]	2020	Thesis
	Zhang and Ramesh [406]	2020	arXiv
	Ignatiev et al. [156]	2020	CP
	Sharma et al. [325]	2021	AIES
	Ezzeldin et al. [113]	2021	arXiv
	Wang et al. [355]	2021	FAccT
	Ozdayi et al. [276]	2021	arXiv
	Zhang et al. [399]	2021	PAKDD
	Perrone et al. [285]	2021	AIES
	Islam et al. [164]	2021	AIES
	Roh et al. [307]	2021	ICLR
	Hort and Sarro [149]	2021	ASE
	Valdivia et al. [345]	2021	Int. J. Intell. Syst.
	Lee et al. [221]	2021	ICML
	Cruz et al. [88]	2021	ICDM
	Roy and Ntoutsi [312]	2022	ECML PKDD
	Wang et al. [351]	2022	arXiv
	Sikdar et al. [329]	2022	FAccT
	Agarwal and Deshpande [13]	2022	FAccT
	Park et al. [279]	2022	WWW
	Djebrouni [100]	2022	Eurosys
	Iosifidis et al. [162]	2022	KAIS
	Short and Mohler [327]	2022	Int. J. Forecast.
	Maheshwari and Perrot [246]	2022	arXiv
	Zhao et al. [408]	2022	KDD
	Tizpaz-Niari et al. [343]	2022	ICSE
	Roy et al. [311]	2022	DS
	Mohammadi et al. [259]	2022	arXiv
	Gao et al. [126]	2022	ICSE
	Huang et al. [154]	2022	Expert Syst. Appl.
	Candelieri et al. [50]	2022	arXiv
	Anahideh et al. [22]	2022	Expert Syst. Appl.
	Rateike et al. [300]	2022	FAccT
	Li et al. [227]	2022	arXiv

(2) choose best split according to $\delta_{accuracy}/\delta_{discrimination}$;
(3) choose best split according to $\delta_{accuracy} + \delta_{discrimination}$.

While constraints and regularization usually utilize group fairness definitions, they have also been applied for achieving individual fairness [108, 129, 177, 205]. Moreover, they can be applied to achieve fairness for multiple sensitive attributes and fairness definitions [194, 194, 195, 209, 278, 342], or extend existing adjustments, such as adding fairness regularization in addition to the L2 norm, which is used to avoid overfitting [188, 191].

4.2.2 *Adversarial Learning.*

Adversarial learning simultaneously trains classification models and their adversaries [91]. While the classification model is trained to predict ground truth values, the adversary is trained to exploit fairness issues. Both models then compete against each other, to improve their performance.

Zhang et al. [391] 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. [39] trained a neural network to predict two outputs: labels and sensitive attributes. While a high overall accuracy is desired, the adversarial setting reduces the ability to predict sensitive information. The network is designed to share layers between the two output, such that only one model is trained [9, 38, 93, 296, 315].

Lahoti et al. [213] 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 [306], to improve individual fairness [380], and to reweigh training data [288]. In particular, Petrović et al. [288] 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) [45, 48, 171, 275, 290, 338, 344] or in an ensemble fashion (i.e., a voting of multiple classification models at the same time) [68, 158, 207, 232, 257, 260, 299, 311].

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. [109] 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. [344] built upon the work of Dwork et al. [109] and incorporated "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. [338] 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 [42, 232, 257, 311, 345]. Practitioners can then choose which fairness-accuracy trade-off best suits their need. For example, Liu and Vicente [232] treated bias mitigation as multi-objective optimization problem that explores fairness-accuracy trade-offs under consideration of multiple fairness metrics. Mishler and Kennedy [257] 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 bias via changing the learning procedure of algorithms or the creation of novel algorithms [107]. Changes have been suggested for a variety of classification models, including Bayesian models [97, 189], Markov Random Fields [406], Neural Networks [152, 251, 296], Decision Trees, bandits [26, 175, 176], boosting [145, 160, 161, 311], Logistic Regression [307]. We outline a selection of publications in the following, to provide insight on techniques applied to different classification models.

Noriega-Campero et al. [270] proposed an active learning framework for training Decision Trees. During 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. [22] used an active learning framework to balance accuracy and fairness by selecting instances to be labelled.

Madras et al. [245] 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 [221].

Martinez et al. [251] 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 [111]. Hu et al. [152] incorporated representation learning into the training procedure of Neural Networks to learn them jointly the classifier.

Hébert-Johnson et al. [145] 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. [144] 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. [199] 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 [61, 62, 88, 89, 149, 164, 285, 343, 345].

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 6), post-processing methods are the least frequently applied of those covered in this survey.

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) [107], which conduct modifications to training data (e.g., relabelling, perturbation and representation learning).

We found only two publications that applied input corrections to testing data, both of which used perturbations. While Adler et al. [10] used perturbation in a post-processing stage, Li et al. [228]

Table 6. Publications on Post-processing bias mitigation methods.

Type	Authors [Ref]	Year	Venue
Input	Adler et al. [10]	2018	KAIS
	Li et al. [228]	2022	ICSE
Output	Pedreschi et al. [281]	2009	SDM
	Kamiran et al. [186]	2012	ICDM
	Fish et al. [119]	2015	FATML
	Fish et al. [120]	2016	SDM
	Liu et al. [230]	2018	arXiv
	Kim et al. [205]	2018	NeurIPS
	Zhang et al. [398]	2018	IJCAI
	Kamiran et al. [187]	2018	J. Inf. Sci.
	Menon and Williamson [256]	2018	FAccT
	Chzhen et al. [77]	2019	NeurIPS
	Chiappa [71]	2019	AAAI
	Iosifidis et al. [158]	2019	Big Data
	Lohia et al. [236]	2019	ICASSP
	Wei et al. [360]	2020	PMLR
	Alabdulmohsin [15]	2020	arXiv
	Alabdulmohsin and Lucic [17]	2021	NeurIPS
	Lohia [235]	2021	arXiv
	Nguyen et al. [269]	2021	J. Inf. Sci.
	Kobayashi and Nakao [207]	2021	DiTTEt
	Jang et al. [166]	2022	AAAI
	Pentyala et al. [283]	2022	arXiv
	Snel and van Otterloo [332]	2022	Com. Soc. Res. J.
	Alghamdi et al. [19]	2022	arXiv
	Mohammadi et al. [259]	2022	arXiv
	Zeng et al. [390]	2022	arXiv
	Zeng et al. [389]	2022	arXiv

Type	Authors [Ref]	Year	Venue
Classifier	Calders and Verwer [48]	2010	DMKD
	Kamiran et al. [185]	2010	ICDM
	Hardt et al. [143]	2016	NeurIPS
	Woodworth et al. [364]	2017	COLT
	Pleiss et al. [290]	2017	NeurIPS
	Gupta et al. [139]	2018	arXiv
	Morina et al. [263]	2019	arXiv
	Noriega-Campero et al. [270]	2019	AIES
	Kanamori and Arimura [192]	2019	JSAI
	Kim et al. [206]	2019	AIES
	Chzhen et al. [78]	2020	NeurIPS
	Chzhen and Schreuder [80]	2020	arxiv
	Savani et al. [321]	2020	NeurIPS
	Awasthi et al. [27]	2020	PMLR
	Kim et al. [202]	2020	ICML
	Jiang et al. [169]	2020	UAI
	Chzhen et al. [79]	2020	NeurIPS
	Du et al. [103]	2021	NeurIPS
	Schreuder and Chzhen [322]	2021	UAI
	Mishler et al. [258]	2021	FAccT
	Mishler and Kennedy [257]	2021	arXiv
	Kanamori and Arimura [193]	2021	JSAI
	Grabowicz et al. [133]	2022	FAccT
	Iosifidis et al. [162]	2022	KAIS
	Mehrabi et al. [254]	2022	TrustNLP
	Zhang et al. [395]	2022	FairWARE
	Wu and He [368]	2022	FAccT
	Marcinkevics et al. [249]	2022	MLHC

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. [321] called intra-processing. A successfully trained classification model is adapted to obtain a fairer one. Such modification have been applied to Naive Bayes [48], Logistic Regression [169], Decision Trees [185, 193, 395], Neural Networks [103, 249, 254, 321] and Regression Models [79].

Hardt et al. [143] 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 [27, 139, 258, 263].

Woodworth et al. [364] 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. [143].

Pleiss et al. [290] 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 prediction. Noriega-Campero et al. [270] followed the calibration approach of Pleiss et al. [290].

Kamiran et al. [185] 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 [193] formulated the modification of branching thresholds for Decision Trees as a mixed integer program.

Kim et al. [206] 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. [281] considered the correction of rule-based classifiers, such as CPAR [378]. 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 [256] proposed a plugin approach for thresholding predictions. To determine the thresholds to use, the class probabilities are estimated using logistic regression.

Kamiran et al. [186, 187] 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. [236] 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. [120] proposed a confidence-based approach which returns a positive label for each prediction above a given threshold. This has also been applied to AdaBoost [119]. Other than using a general threshold for all instances, group dependent thresholds can be used [15, 77, 158, 166, 207, 283, 389, 390].

Chiappa [71] 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 proposed 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 7 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 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 [46, 163, 184, 335, 416]. 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 [128]. Moreover, there are three publications that addressed bias mitigation at each stage: pre-processing, in-processing and post-processing [48, 139, 158].

Calders and Verwer [48] 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

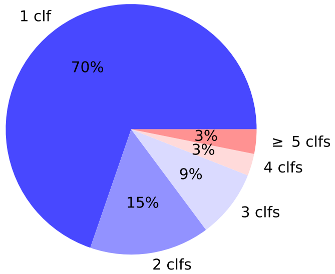


Fig. 4. Number of classification models (clf) used for evaluation.

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. [139] 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. Lastly, they followed the approach of Hardt et al. [143] to debias and existing classifier by adding an additional variable to the prediction problem (see Section 4.3.2).

Iosifidis et al. [158] 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.

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 8 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 second most frequently used classification models are Neural Networks (NNs). 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 4 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

Table 7. Publications with multiple bias mitigation methods. "X" indicates that the publication applies a bias mitigation approach of the corresponding category.

	Processing Method		
Authors	Pre	In	Post
Sun et al. [335]	x x	x	
Calders et al. [46]	x x		
Žliobaite et al. [416]	x x		
Hajian and Domingo-Ferrer [142]	x x		
Kamiran and Calders [184]	x x		
Iosifidis et al. [163]	x x		
Chakraborty et al. [60]	x x		
Oneto et al. [275]	x	x x	
Calders and Verwer [48]	x	x	x
Gupta et al. [139]	x	x	x
Iosifidis et al. [158]	x	x	x
Pérez-Suay et al. [284]	x	x	
Komiyama and Shimao [208]	x	x	
Kilbertus et al. [201]	x	x	
Grgić-Hlača et al. [138]	x	x	
Madras et al. [244]	x	x	
Xu et al. [373]	x	x	
Abay et al. [7]	x	x	
Hu et al. [152]	x	x	
Chakraborty et al. [61]	x	x	
Chuang and Mroueh [76]	x	x	
Zhang et al. [399]	x	x	
Grari et al. [136]	x	x	
Du and Wu [104]	x	x	
Amend and Spurlock [21]	x	x	
Cruz et al. [88]	x	x	
Chen et al. [64]	x	x	
Liang et al. [229]	x	x	
Agarwal and Deshpande [13]	x	x	
Chen et al. [68]	x	x	
Wu et al. [366]	x	x	
Rateike et al. [300]	x	x	
Kim and Cho [204]	x	x	
Suriyakumar et al. [338]	x	x	
Zhang et al. [398]	x		x
Wei et al. [360]	x		x
Pentyala et al. [283]	x		x
Li et al. [228]	x		x
Mishler and Kennedy [257]		x x x	x
Quadrianto and Sharmanska [293]		x x	
Agarwal et al. [11]		x x	
Gillen et al. [129]		x x	
Kearns et al. [195]		x x	
Goel et al. [130]		x x	
Beutel et al. [38]		x x	
Kilbertus et al. [200]		x x	
Liu et al. [233]		x x	
Kamani [180]		x x	
Perrone et al. [285]		x x	
Grari et al. [135]		x x	
Ranzato et al. [299]		x x	
Park et al. [279]		x x	
Wang et al. [354]		x x	
Zhao et al. [408]		x x	
Roy et al. [311]		x x	
Boulitsakis-Logothetis [45]		x x	
Kamiran et al. [185]		x	x
Fish et al. [119]		x	x
Woodworth et al. [364]		x	x
Pleiss et al. [290]		x	x
Kim et al. [205]		x	x
Chiappa [71]		x	x
Noriega-Campero et al. [270]		x	x
Chzhen and Schreuder [80]		x	x
Kim et al. [202]		x	x
Jiang et al. [169]		x	x
Chzhen et al. [79]		x	x
Kobayashi and Nakao [207]		x	x
Iosifidis et al. [162]		x	x
Mohammadi et al. [259]		x	x

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.

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.

Table 8. 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.

		Method		
Model	Unique	Pre	In	Post
Logistic Regression	140	58	80	19
Neural Network	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
Nearest Neighbor	13	7	2	5
AdaBoost	8	1	5	4
XGBoost	8	1	6	1
Causal	7	2	6	1
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

Following this procedure, we gathered a total of 83 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 the creation of such synthetic data.

For further details on datasets, we refer to Le Quy et al. [218] who surveyed 15 datasets and provided detailed information on the features and dataset characteristics. Additionally, Kuhlman et al. [211] 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). Fairness datasets for a variety of domains (e.g., health, linguistics, social sciences, computer vision) can be found in the web app by Fabris et al. [114].[4]

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. For this purpose, we consider **324 of the 341 publications**, as only these 324 publications perform empirical experiments. The remaining publications do not present any empirical experiment and thus do not consider any dataset.

[4] http://fairnessdata.dei.unipd.it/

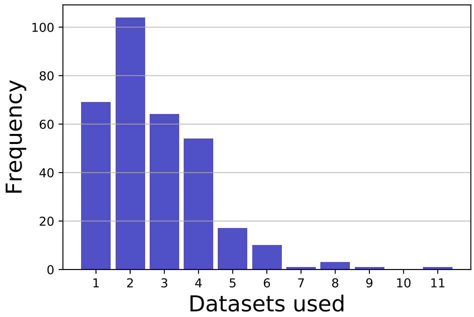


Fig. 5. Number of datasets used per publication.

Among the 83 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 only once during 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 9 list the frequency of the remaining 22 datasets (used in three or more publications). A list of all datasets can be found in our online repository [25]. 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. [98] 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. However, these two datasets have recently received criticism for their use as benchmark datasets and suitability as real-world datasets. For instance, the Adult dataset applies a binary label to determine whether an individual has an income above 50,000 USD. Ding et al. [98] showed that the fairness of ML models and bias mitigation methods is depending on the income threshold, thereby potentially limiting the external validity of the Adult dataset for benchmarking. Bao et al. [32] addressed the use of Risk assessment instrument (RAI) datasets, in particular the COMPAS dataset, for benchmarking ML fairness. They outlined that the use of such datasets should consider domain context, rather than using them as a generic example to show the real-world performance of bias mitigation methods.

5.2 Dataset Frequency

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 [211]. Figure 5 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. Overall, 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. [59]), and 11 datasets (Do et al. [101]) respectively. In accordance with existing work, new publications should evaluate their bias mitigation methods on three datasets, and if possible more.

Table 9. Frequency of widely used datasets (i.e., used in at least three publications).

Dataset Name	Frequency	Percentage
Adult [106]	249	77%
COMPAS [24]	166	51%
German [106]	97	30%
Communities and Crime [302]	42	13%
Bank [264]	38	12%
Law School [362]	33	10%
Default [377]	24	7%
Dutch Census [1]	16	5%
Health [3]	14	4%
MEPS [2]	14	4%
Drug [116]	9	3%
Student [83]	8	2%
Heart disease [106]	7	2%
National Longitudinal Survey of Youth [6]	6	2%
SQF [4]	5	2%
Arrhythmia [106]	5	2%
Wine [82]	4	1%
Ricci [337]	4	1%
University Anonymous (UNIV)	3	1%
Home credit [5]	3	1%
ACS [98]	3	1%
MIMICIII [173]	3	1%

Hereby it can be of interested to consider a diverse range of datasets based on application domains, dimensionality or protected attributes [218].

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 [109, 200, 244].

The created datasets range from hundreds of data points [97, 144, 215, 273] to 100,000 and above [95, 147, 163, 386]. While the sampling procedures are well described, some publications do not state the dataset size used for experiments [29, 74, 133, 135, 202, 247, 254, 391].

As exemplary data creation procedure, we briefly outline the data generation approach applied by Zafar et al. [385], as it is the most frequently adapted approach by other publications [199, 202, 232, 279, 306–308, 384]. In particular, Zafar et al. [385] 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 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 [122] therefore improving the conclusion validity of a study [292]. 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 [50, 86, 154, 196, 215, 224, 274, 329, 374]).

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. [84] 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 [130].

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) and 70%-30% (35 times) followed by 67%-33% (16 times), 50%-50% (11 times), 60%-40% (5 times), and 75%-25% (5 times). In addition to these regular sized datas-plits, there are 23 publication which divide the data into very "specific" splits. For example, Quadrianto et al. [294] divided the Adult dataset into $28,222$ training, $15,000$ and $2,000$ validation instance. Another example is the work by Liu and Vicente [232], 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 in the work by da Cruz [89], 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 [34].

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 [257] 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 [161, 163, 327, 399, 400], federated learning [7, 113, 151, 283, 291], multi-source [157], sequential [20, 300, 407, 408] 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. In particular, this section focuses on metrics used for measuring bias, rather than general notions of fairness such as *Fairness through Unawareness* (i.e., not using the protected attribute).

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 Caton and Haas [51], and Verma and Rubin [348].

Table 10. Popular fairness metrics. At least one metric for each category is provided.

Name	Section	#	Description
Statistical Parity Difference	6.2	136	Difference of positive predictions per group
Equality of Opportunity	6.3	91	Equal TPR per population groups
Disparate Impact, P-rule	6.2	60	Ratio of positive predictions per group
Equalized Odds	6.3	51	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

Overall we categorize the metrics used in the 341 publications in six categories, which are defined based on labels in dataset; predicted outcome; predicted and actual outcomes; predicted probabilities and actual outcome; similarity; causal reasoning.

In the following, we provide information on how these metric types have been used. In total, we found 109 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 [16, 53, 56, 325, 390], and multi-class predictions [16, 19].

While some of the categories only contain few different metrics (definitions based on labels in dataset, on predicted probabilities and actual outcome, and on similarity all have 13 or fewer different metrics); *definitions based on predicted outcome* have 22, *definitions based on predicted and actual outcomes* have 31, and *definitions based on Causal Reasoning* 27 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 10, while listing at least one metric per category. For detailed explanations of fairness metrics, we refer to Verma and Rubin [348].

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 [138]) and quality of representations [253, 319, 333] (see Section 4.1.4).

Notations. To provide equations of fairness metrics, we use the following notation:

- S: sensitive attribute to divide populations in two groups (s_1, s_2).
- y: Ground truth label.
- $\hat{y}$: Predicted label (or probability, Section 6.4).
- Pr: Probability.
- D: Dataset, with N instances.

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 [35]. 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, 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 (MD), slift or elift [263]. Hereby, MD is the most popular, used in 6 publications.

$$MD = Pr(y = 1|S = s_1) - Pr(y = 1|S = s_2)$$
$$elift = e^{-\epsilon} \leq \frac{Pr(y = 1|S = s)}{Pr(y = 1)} \leq e^{\epsilon}, \forall s \in S$$
$$slift = e^{-\epsilon} \leq \frac{Pr(y = 1|S = s)}{Pr(y = 1|S = s')} \leq e^{\epsilon}, \forall s, s' \in S$$

elift and slift are parameterized by ϵ, which allows for an easy comparison of bias between different classification models, by contrasting the magnitude of their ϵ values. Perfect fairness is achieved by $\epsilon = 0$.

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.

$$\text{Statisitcal Parity Difference (SPD)} = Pr(\hat{y} = 1|S = s_1) - Pr(\hat{y} = 1|S = s_2)$$
$$\text{Disparate Impact (DI)} = \frac{Pr(\hat{y} = 1|S = s_1)}{Pr(\hat{y} = 1|S = s_2)}$$
$$\text{P-rule} = min\left(\frac{Pr(\hat{y} = 1|S = s_1)}{Pr(\hat{y} = 1|S = s_2)}, \frac{Pr(\hat{y} = 1|S = s_2)}{Pr(\hat{y} = 1|S = s_1)}\right)$$

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 [93, 288, 289, 296]. 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 [415] 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.

Frequently, metrics based on predicted and actual outcomes are computed from combinations of confusion matrix measures (i.e., True Positives (TP), False Positives (FP), False Negatives (FN), True Negatives (TN)):

$$\begin{aligned}
\text{True Positve Rate (TPR)} &= \frac{TP}{TP+FN}\\
\text{False Positve Rate (FPR)} &= \frac{FP}{FP+TN}\\
\text{False Negative Rate (FNR)} &= \frac{FN}{FN+TP}\\
\text{True Negative Rate (TNR)} &= \frac{TN}{TN+FP}\\
\text{Positive Predictive Rate (PPR)} &= \frac{TP}{TP+FP}\\
\text{Negative Predictive Rate (NPR)} &= \frac{TN}{TN+FN}\\
\text{False Discovery Rate (FDR)} &= \frac{FP}{TP+FP}
\end{aligned}$$

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. An average score of TPR and FPR is provided by the *Average Odds Difference*.

$$\begin{aligned}
\text{Equality of Opportunity} &= TPR_{S=s_1} - TPR_{S=s_2}\\
\text{Equalized Odds} &= (FPR_{S=s_1} - FPR_{S=s_2}) + (TPR_{S=s_1} - TPR_{S=s_2})\\
\text{Average Odds} &= \frac{1}{2}((FPR_{S=s_1} - FPR_{S=s_2}) + (TPR_{S=s_1} - TPR_{S=s_2}))
\end{aligned}$$

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 [316, 318] and one can compare populations groups with regards to performance metrics, such as precision, recall, F1 and Area Under Curve.

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. [169] 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):

$$\text{Strong Pairwise Demographic Parity (SPDD)} = \mathbb{E}_{\tau\sim U(\Omega))}|Pr(\hat{y} > \tau|S = s_1) - Pr(\hat{y} > \tau|S = s_2)|$$

Here, $\mathbb{E}_{\tau\sim U(\Omega))}$ denotes the expectation over all possible thresholds τ, uniformly sampled from all possible prediction outcomes $U(\Omega)$. Chzhen et al. [79] also varied thresholds, to compute the Kolmogorov-Smirnov distance. Heidari et al. [146] measured fairness based on positive and negative residual differences. Agarwal et al. [12] computed a Bounded Group Loss (BGL) to minimize the worst loss of any group, according to least squares.

Another notion of fairness based on predicted probabilities and actual outcomes is calibration [290]. Calibration describes a scenario where predicted probabilities have a semantic meaning, for example if 100 individuals receive a prediction of 0.75, then 75 of them should have a positive label (i.e., a label of 1). Zhang and Weiss [402, 403] proposed the use of a related metric with *fair calibration* (FC). FC first sorts predicted probabilities for each subgroup and divides them in 10 equally sized bins (e.g., 100 instances would result in 10 bins of 10 individuals). It is then evaluated whether the 10 bins of each population group are calibrated, and in a second stage whether differences between predictions and actual outcomes are consistent across population groups. FC then generates a binary result, whether the model is fairly calibrated or not.

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) [388]. *Consistency* compares the prediction of an individual with the k-nearest-neighbors according to the input space [388]. Luong et al. [241] also utilized k-nearest-neighbors, to investigate the difference in predictions for different values of k.

$$\text{Consistency} = 1 - \frac{1}{Nk} \sum_{n} |\hat{y}_n - \sum_{j \in kNN(x_n)} \hat{y}_j|$$

Similarities between individuals have been computed according to ℓ_∞-distance [313], and euclidean distance with weights for features [388]. Individuals have also been treated as similar if they have equal labels [36], are equal except for sensitive features or based on predicted labels [347]. If similarity of individuals is determined solely by differences in sensitive features, one is speaking of “causal discrimination” [236, 414].[5]

In contrast to determining similarity computationally, Jung et al. [177] allowed stakeholders to judge whether two individuals should receive the same treatment.

Moreover, Ranzato et al. [299] considered four types of similarity relations (Noise, Cat, Noise-Cat, conditional-attribute), when dealing with numerical and categorical features. Verma et al. [347] considered two types of similarities: input space (identical on non-sensitive features), output space (identical prediction). Lahoti et al. [215] 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 [348].

For example, Counterfactual fairness states that a causal graph is fair, if the prediction does not depend on descendants of the protected attribute [212]. 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 [397]. 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.

[5] Some publications refer to this as “Counterfactual fairness’ [260, 380, 381], but we follow the guidelines of Verma and Rubin [348] and treat counterfactual fairness as a Causal metric.

Table 11. 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.

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

Direct and indirect discrimination can be modelled based on the causal effect along paths taken in causal graphs [397]. To measure indirect discrimination, Prejudice Index (PI) or Normalized Prejudice Index (NPI) haven been applied four times [188]. NPI quantifies the mutual information between protected attributes and predictions.

$$PI = \sum_{y,s \in D} Pr(y,s) ln \frac{Pr(y,s)}{Pr(s)Pr(y)}$$

$$NPI = PI/(\sqrt{H(Y)H(S)}))$$

Here, $H(X)$ is defined as the entropy function $-\sum_{x \in D} Pr(x) ln Pr(x)$.

Mutual information has also been used to determine the fairness of representations [265, 333]. 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 nine 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 (i.e., out of the 341 publications, 324 perform experiments, among which 308 perform 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. We use the term "baseline" to describe simple methods for benchmarking, that can be applied as sanity checks to determine whether a bias mitigation methods is effective. Unlike methods presented in Section 7.2 and Section 7.3, these are not based on published methods.

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*. 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 [46, 282], as the remaining features are often correlated with the protected attribute. To combat this risk, Kamiran et al. [185] 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) [202, 259, 265, 357]. This serves as a fairness-aware baseline, as every individual and population group receive the same treatment [150].

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 11 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. 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 6 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 [184], constraints [382, 385] and adversarial learning [391] for in-processing, and classifier modification for post-processing [143].

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 12 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. A total of 23 and 27 publications for pre-processing and in-processing methods, respectively, benchmark against fairness-unaware methods.

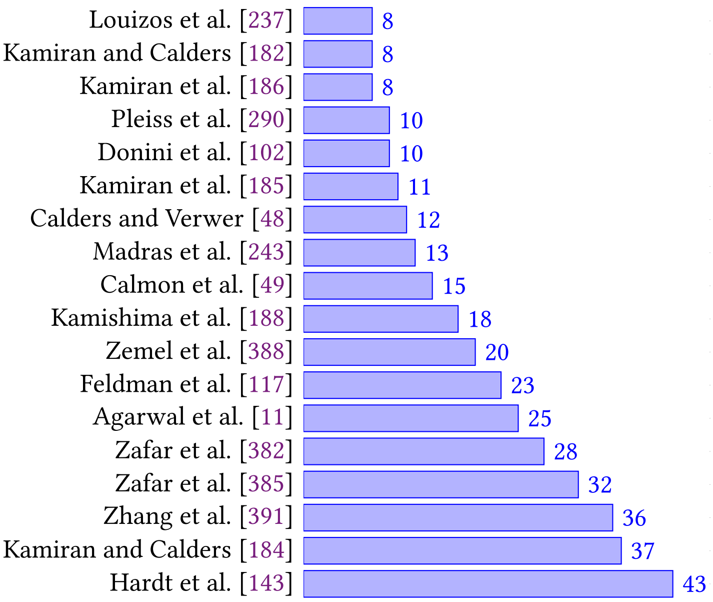


Fig. 6. Most frequently benchmarked publications. For each publication, the number of times it has been used for benchmarking is shown.

Table 12. Publications that benchmark against at least one fairness-unaware method.

Type	Category	Section	References
Pre	Sampling	4.1.2	Abusitta et al. [8], Celis et al. [55], Cruz et al. [88], Xu et al. [373] Du and Wu [104], Roh et al. [308], Xu et al. [370], Yan et al. [374] Dablain et al. [90], Pentyala et al. [283], Zhang et al. [399]
	Representation	4.1.4	Creager et al. [87], Gupta et al. [140], Louizos et al. [237], Salazar et al. [316] Balunović et al. [30], Galhotra et al. [125], Oh et al. [272], Qi et al. [291] Jaiswal et al. [165], Lahoti et al. [214], Sarhan et al. [320], Shui et al. [328]
In	Regularization	4.2.1	Jiang et al. [170], Liu et al. [233], Wang et al. [354], Zhang and Weiss [402, 403]
	Constraints	4.2.1	Ding et al. [99], Du and Wu [104], Zhang et al. [392], Zhao et al. [407] Fukuchi et al. [123], Narasimhan [268], Wang et al. [354], Zhao et al. [408]
	Adversarial	4.2.2	Lahoti et al. [213], Roh et al. [306], Sadeghi et al. [315], Xu et al. [373] Rezaei et al. [304], Yazdani-Jahromi et al. [376]
	Adjusted	4.2.4	Cruz et al. [88], Iosifidis and Ntoutsi [161], Liu et al. [233], Luo et al. [240] Candelieri et al. [50], Sharma et al. [325], Wang et al. [351], Zhang et al. [399] Lee et al. [221], Maheshwari and Perrot [246], Zhao et al. [408]
Post	Input	4.3.1	Adler et al. [10]
	Classifier	4.3.2	Mehrabi et al. [254], Wu and He [368]
	Output	4.3.3	Alabdulmohsin and Lucic [17], Kamiran et al. [187], Pentyala et al. [283]

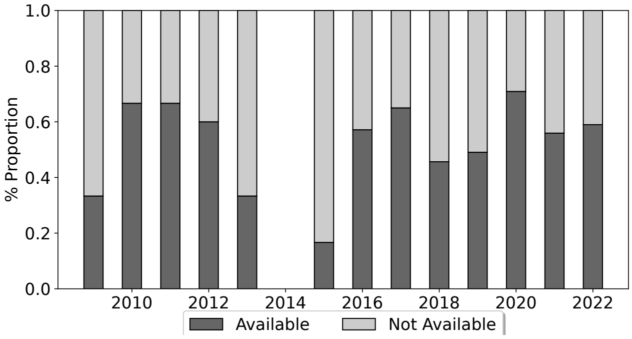


Fig. 7. Proportion of publications that shared implementation source code, per year.

7.4 Source Code Availability

To investigate whether existing work allows for reproducibility of the results and ease of use for benchmarking, we reviewed whether the 341 surveyed publications shared source code. Specifically, we have collected links to implementations from the publications directly. If no link was available, we performed a google search to check for resources we missed.[6] With this additional search, we were able to find 64 implementations. Overall, we found 192 publications with available source code (56% of the 341 publications).

Figure 7 illustrates the proportion of publications with code available per year. Early years (2009-2016) show a high variation in the proportion of publications with source code available, ranging from 17% to 67%. Such a variation is caused by the small number of publications. In 2018 and 2019, the proportion of publications with shared source code is below 50%, 46% and 49% respectively. The most recent years showed an increase in shared implementations, with the maximum achieved in 2020 with 71% of the publications to share source code.

Moreover, we examined existing surveys for frameworks which provide implementations of bias mitigation methods [51, 67, 107, 222, 255, 287, 334], and found three frameworks that do so: Themis-ML [31], AIF 360 [35], Fairlearn [41]. In total, Themis-ML [31] implements bias mitigation from three publications, Fairlearn [41] implements four methods and AIF 360 [35] implements 13 methods.[7]

While our focus lies on the sharing and reuse of bias mitigation methods, datasets are also an important resource to share, to allow for a reproducibility of conducted experiments. Many datasets are already shared, however some datasets are proprietary and cannot be shared publicly. Where available, we provide links to datasets and source code implementations in our online repository [25].

8 CHALLENGES AND OPPORTUNITIES

This section provides further discussion and insights on the surveyed publications. We outline several challenges, based on the current literature, as well as discuss research opportunities for the creation and evaluation of new bias mitigation methods.

[6] For each publications, we searched for “*paper title*” and “paper title github” and checked the first page of search results for links to external resources.

[7] 1st of March 2023

8.1 Challenges

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

8.1.1 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. Consolidating or selecting a fixed set of metrics to use is still an open challenge [90, 127, 255], as can be seen by the 109 different fairness metrics obtained in Section 6.

While consolidating 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 [54, 154, 184, 260, 263] 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 [11, 51, 271, 286], 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) [48, 183, 184, 388].

8.1.2 Fairness Guarantees. Guarantees are of particular importance when dealing with domains that fall under legislation and regulatory controls [117, 188]. Thereby, 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. Fairness guarantees can help in this situation, by providing performance guarantees with regards to a specific fairness metric and bound the degree of bias [52, 176]. In particular, Dunkelau and Leuschel [107] 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 respect to a given fairness metric and 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 [172, 188, 294, 364].

8.1.3 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 [148].

Moreover, it is not guaranteed that the annotation process of the training data is bias free [143]. If possible an unbiased data collection should be enforced [293]. Other options are the debiasing of ground truth labels [379, 414] or use of expert opinions to annotate data [103]. If feasible, more data can be collected [65, 172], which is difficult from a research perspective, as commonly, existing and public datasets are used without the chance to manually collect new samples.

Besides, the variety of protected attributes addressed in previous experiments, as found by Kuhlman et al. [211], 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. [148]. However, an interviewee questioned whether such a fairness intervention is fair, as the targeting of subgroups for additional data collection may be a biased procedure.

8.1.4 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 [141] 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 [148], 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" [35]). Therefore, Martinez et al. [251] 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 [220]. 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) [23, 254]. Nonetheless, Sylvester and Raff [339] argue that considering fairness criteria while developing ML models is better than considering none, even if the metric is not optimal.

Sharma et al. [326] 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 [35] and Fairlearn [41] can be used.

8.1.5 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 [50, 61, 62, 89, 149, 210, 355, 374].

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 [194, 385], ranking [153, 188, 385] and clustering [188].

8.2 Research Opportunities

In the course of this survey, we have collected 341 publications with regards to various approaches for bias mitigation methods. This collection helps us to understand which approaches have already been applied and allows us to outline some aspects that appear underexplored and provide opportunities for future research.

Firstly, from the 341 publications we collected, it can be seen that in-processing methods are the most widely explored methods. There are almost twice as many publications with in-processing methods than pre-processing, and nearly four times as many in-processing methods than post-processing methods. Therefore, addressing post-processing bias mitigation method seems promising in contrast to the other two method types. In particular the modification of inputs in a post-processing stage has only been considered by two publications (Section 4.3.1) [10, 228]. However, this type of bias mitigation method could be further investigated without considerable effort by developing new methods, simply by applying existing pre-processing methods (Section 4.1) to the testing data.

Generally speaking, pre- and post-processing methods are classifier-agnostic and can be evaluated on a variety of classification models without modification to the underlying algorithm. Nonetheless, Bandits have been investigated with neither of these two method types, only by in-processing methods [129, 175, 176].

Moreover, the combination of pre- and post-processing methods has only been addressed four times [228, 283, 360, 398]. The number of classification models considered by these four publications range from 1 to 3. This is a promising combination of approaches, as one can perform experiments with bias mitigation methods at two different stages (i.e., before and after training) on various classification models and thereby collect extensive empirical evidence for fairness improvements. Additionally, we found several publications that applied multiple bias mitigation methods of the same type (e.g., two pre-processing methods). Six of these applied multiple pre-processing methods and 19 applied multiple in-processing methods (Table 7). However, we found no publication that applied multiple post-processing methods.

Lastly, our data collection shows that there exist a multitude of datasets and metrics, which can enable a rigorous evaluation of novel bias mitigation methods. For one, bias mitigation methods can be evaluated on up to 81 datasets, whereas bias mitigation methods evaluated on three datasets exceed the average of 2.7 datasets used for evaluation. When applying bias mitigation methods to a dataset, it is important to mention the protected attributes considered and potential criticisms that could impact the ability to make claims about applicability for real world systems [32, 98].

The 109 metrics are divided in six categories. Thereby, bias mitigation method can be evaluated by multiple metrics of the same category, or metrics from multiple categories. In addition to using fairness metrics to evaluate the performance of bias mitigation methods, performance metrics, such as accuracy, can be used to determine the fairness-accuracy trade-off achieved when applying bias mitigation methods. To ensure the competitiveness of results, methods must always be benchmarked against baselines as well as previous existing relevant methods, especially when their implementation is made publicly available (our survey highlights that 192 studies provided source code implementations, and as such they could be used as a benchmark for future proposals).

9 CURRENT BEST PRACTICES / RECOMMENDATIONS

In this section, we would like to outline current practices for the empirical evaluation of bias mitigation methods, that we have observed from the 341 publications. However, we note that increasing the comprehensiveness of the empirical evaluation is always positive to support the validity of results (e.g., applying bias mitigation methods to a higher number of datasets, or using more metrics for evaluation). Our recommendations, which will allow new experiments to be in line with prior experiments conducted, are as follows: **1.** Check existing approaches, to confirm the novelty of the bias mitigation method under evaluation.

2. Apply your bias mitigation method to at least three datasets, taking diversity and criticism into account when making claims about real world impact.

3. State the protected attributes for each dataset.

4. Evaluate your bias mitigation method on at least two fairness metrics, as well as an performance metric (e.g., accuracy). We suggest using different metric types to reduce the correlation of individual fairness metrics.

5. Benchmark at least against the original model and consider similar, existing bias mitigation methods as well.

6. Apply your bias mitigation method to multiple classification models, in particular when proposing pre- or post-processing methods. Logistic regression and neural networks are frequently used.

7. Try to repeat experiments at least 10 times for standard training splits (e.g. 70% or pre-defined data-splits).
8. Share code and numerical results, in particular when results are presented in bar charts.

10 CONCLUSION

In this literature survey, we 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). 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 (109 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. Among the collected publications, we found 56% (192 out of 341) that make source code available, which supports the replicability of results and benchmarking efforts. Moreover, we have found three frameworks to implement and make available existing bias mitigation methods [31, 35, 41].

Lastly, we list current opportunities and challenges that have been discerned from 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 support the development of new methods.

ACKNOWLEDGEMENTS

Zhenpeng Chen, Mark Harman and Federica Sarro are supported by the ERC grant no. 741278 (EPIC). Jie M. Zhang is partially supported by the UKRI Trustworthy Autonomous Systems Node in Verifiability, with Grant Award Reference EP/V026801/2. Max Hort is supported through the ERCIM 'Alain Bensoussan' Fellowship Programme. We would like to thank the members of the community who kindly provided comments and feedback on an earlier version of this article.

REFERENCES

[1] 2001. Dutch Central Bureau for Statistics Volkstelling. http://easy.dans.knaw.nl/dms. Retrieved on June 12, 2022.

[2] 2016. Medical Expenditure Panel Survey dataset. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192. Retrieved on June 12, 2022.

[3] 2017. The Heritage Health Prize dataset. https://www.kaggle.com/c/hhp. Retrieved on June 12, 2022.

[4] 2017. Stop, Question and Frisk dataset. http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page. Retrieved on June 12, 2022.

[5] 2018. Home Credit Default Risk. https://www.kaggle.com/c/home-credit-default-risk. Retrieved on June 12, 2022.

[6] 2019. National longitudinal surveys of youth data set. www.bls.gov/nls/. Retrieved on June 12, 2022.

[7] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. 2020. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447* (2020).

[8] Adel Abusitta, Esma Aïmeur, and Omar Abdel Wahab. 2019. Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems. *arXiv preprint arXiv:1905.09972* (2019).

[9] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2412–2420.

[10] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54, 1 (2018), 95–122.

[11] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[12] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[13] Sushant Agarwal and Amit Deshpande. 2022. On the Power of Randomization in Fair Classification and Representation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1542–1551. https://doi.org/10.1145/3531146.3533209

[14] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1418–1426.

[15] Ibrahim Alabdulmohsin. 2020. Fair Classification via Unconstrained Optimization. *arXiv preprint arXiv:2005.14621* (2020).

[16] Ibrahim Alabdulmohsin, Jessica Schrouff, and Oluwasanmi Koyejo. 2022. A Reduction to Binary Approach for Debiasing Multiclass Datasets. *arXiv preprint arXiv:2205.15860* (2022).

[17] Ibrahim M Alabdulmohsin and Mario Lucic. 2021. A near-optimal algorithm for debiasing trained machine learning models. *Advances in Neural Information Processing Systems* 34 (2021), 8072–8084.

[18] Daniel Alabi, Nicole Immorlica, and Adam Kalai. 2018. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory*. PMLR, 2043–2066.

[19] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, P Winston Michalak, Shahab Asoodeh, and Flavio P Calmon. 2022. Beyond Adult and COMPAS: Fairness in Multi-Class Prediction. *arXiv preprint arXiv:2206.07801* (2022).

[20] Abdulaziz A. Almuzaini, Chidansh A. Bhatt, David M. Pennock, and Vivek K. Singh. 2022. ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1552–1560. https://doi.org/10.1145/3531146.3533211

[21] Jack J Amend and Scott Spurlock. 2021. Improving machine learning fairness with sampling and adversarial learning. *Journal of Computing Sciences in Colleges* 36, 5 (2021), 14–23.

[22] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. 2022. Fair active learning. *Expert Systems with Applications* 199 (2022), 116981. https://doi.org/10.1016/j.eswa.2022.116981

[23] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*. PMLR, 314–323.

[24] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2016).

[25] Online Appendix. 2022. Online Appendix: Survey Results. https://docs.google.com/spreadsheets/d/1kOmbKLMiFgHRSXvgM-O8OW4YKeDIGN0cPPeCGQOMnnA/edit?usp=sharing

[26] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2 (2002), 235–256.

[27] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. 2020. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1770–1780.

[28] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. 2020. R\'enyi Fair Inference. In *8th International Conference on Learning Representations, ICLR 2020*.

[29] Ananth Balashankar, Alyssa Lees, Chris Welty, and Lakshminarayanan Subramanian. 2019. What is fair? exploring pareto-efficiency for fairness constrained classifiers. *arXiv preprint arXiv:1910.14120* (2019).

[30] Mislav Balunović, Anian Ruoss, and Martin Vechev. 2022. Fair Normalizing Flows. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BrFIKuxrZE

[31] Niels Bantilan. 2018. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services* 36, 1 (2018), 15–30.

[32] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498* (2021).

[33] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[34] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).

[35] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[36] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[37] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[38] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[39] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[40] Peter J Bickel, Eugene A Hammel, and J William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404.

[41] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[42] William Blanzeisky and Pádraig Cunningham. 2022. Using Pareto simulated annealing to address algorithmic bias in machine learning. *The Knowledge Engineering Review* 37 (2022), e5. https://doi.org/10.1017/S0269888922000029

[43] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[44] Ruth G Blumrosen. 1978. Wage discrimination, job segregation, and the title vii of the civil rights act of 1964. *U. Mich. JL Reform* 12 (1978), 397.

[45] Stelios Boulitsakis-Logothetis. 2022. Fairness-Aware Naive Bayes Classifier for Data with Multiple Sensitive Features. *arXiv preprint arXiv:2202.11499* (2022).

[46] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[47] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*. IEEE, 71–80.

[48] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[49] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[50] Antonio Candelieri, Andrea Ponti, and Francesco Archetti. 2022. Fair and Green Hyperparameter Optimization via Multi-objective and Multiple Information Source Bayesian Optimization. *arXiv preprint arXiv:2205.08835* (2022).

[51] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[52] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.

[53] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*. PMLR, 1349–1361.

[54] L Elisa Celis and Vijay Keswani. 2019. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443* (2019).

[55] L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. 2020. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International Conference on Machine Learning*. PMLR, 1349–1359.

[56] L Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. 2021. Fair classification with adversarial perturbations. *Advances in Neural Information Processing Systems* 34 (2021), 8158–8171.

[57] Mattia Cerrato, Alesia Vallenas Coronel, Marius Köppel, Alexander Segner, Roberto Esposito, and Stefan Kramer. 2022. Fair Interpretable Representation Learning with Correction Vectors. *arXiv preprint arXiv:2202.03078* (2022).

[58] Junyi Chai and Xiaoqian Wang. 2022. Fairness with Adaptive Weights. In *International Conference on Machine Learning*. PMLR, 2853–2866.

[59] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. 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* (Athens, Greece) *(ESEC/FSE 2021)*. Association for Computing Machinery, New York, NY, USA, 429–440. https://doi.org/10.1145/3468264.3468537

[60] Joymallya Chakraborty, Suvodeep Majumder, and Huy Tu. 2022. Fair-SSL: Building fair ML Software with less data. In *International Workshop on Equitable Data and Technology (FairWare '22)*.

[61] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. *Fairway: A Way to Build Fair ML Software*. Association for Computing Machinery, New York, NY, USA, 654–665. https://doi.org/10.1145/3368089.3409697

[62] Joymallya Chakraborty, Tianpei Xia, Fahmid M Fahid, and Tim Menzies. 2019. Software engineering for fairness: A case study with hyperparameter optimization. *arXiv preprint arXiv:1905.05786* (2019).

[63] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[64] Canyu Chen, Yueqing Liang, Xiongxiao Xu, Shangyu Xie, Yuan Hong, and Kai Shu. 2022. On Fair Classification with Mostly Private Sensitive Attributes. *arXiv preprint arXiv:2207.08336* (2022).

[65] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems* 31 (2018).

[66] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.

[67] Zhenpeng Chen, Jie M Zhang, Max Hort, Federica Sarro, and Mark Harman. 2022. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *arXiv e-prints* (2022), arXiv–2207.

[68] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2022. 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.*

[69] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2023. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Trans. Softw. Eng. Methodol.* 32, 4 (2023), 106:1–106:30.

[70] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. An Empirical Study on Fairness Improvement with Multiple Protected Attributes. *CoRR* abs/2308.01923 (2023).

[71] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

[72] Silvia Chiappa and William S Isaac. 2018. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.

[73] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. 2020. A fair classifier using kernel density estimation. *Advances in Neural Information Processing Systems* 33 (2020), 15088–15099.

[74] YooJung Choi, Meihua Dang, and Guy Van den Broeck. 2021. Group Fairness by Probabilistic Modeling with Latent Fair Decisions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 12051–12059.

[75] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[76] Ching-Yao Chuang and Youssef Mroueh. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=DNl5s5BXeBn

[77] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2019. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems* 32 (2019).

[78] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2020. Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems* 33 (2020), 19137–19148.

[79] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2020. Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems* 33 (2020), 7321–7331.

[80] Evgenii Chzhen and Nicolas Schreuder. 2020. A minimax framework for quantifying risk-fairness trade-off in regression. *arXiv preprint arXiv:2007.14265* (2020).

[81] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[82] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems* 47, 4 (2009), 547–553.

[83] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. (2008).

[84] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*. PMLR, 1397–1405.

[85] Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. 2019. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *J. Mach. Learn. Res.* 20, 172 (2019), 1–59.

[86] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. 2019. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*. PMLR, 300–332.

[87] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*. PMLR, 1436–1445.

[88] André F Cruz, Pedro Saleiro, Catarina Belém, Carlos Soares, and Pedro Bizarro. 2021. Promoting Fairness through Hyperparameter Optimization. In *2021 IEEE International Conference on Data Mining (ICDM)*. 1036–1041. https://doi.org/10.1109/ICDM51629.2021.00119

[89] André Miguel Ferreira da Cruz. 2020. Fairness-Aware Hyperparameter Optimization: An Application to Fraud Detection. (2020).

[90] Damien Dablain, Bartosz Krawczyk, and Nitesh Chawla. 2022. Towards A Holistic View of Bias in Machine Learning: Bridging Algorithmic Fairness and Imbalanced Learning. *arXiv preprint arXiv:2207.06084* (2022).

[91] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. 2004. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 99–108.

[92] Jeffrey Dastin. 2018. *Amazon scraps secret AI recruiting tool that showed bias against women*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[93] Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt. 2020. Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning. In *Bias and Fairness in AI (BIAS 2020)*.

[94] Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. 2022. FIFA: Making Fairness More Generalizable in Classifiers Trained on Imbalanced Data. *arXiv preprint arXiv:2206.02792* (2022).

[95] Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. 2020. Counterfactual fairness: removing direct effects through regularization. *arXiv preprint arXiv:2002.10774* (2020).

[96] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2022. Multiaccurate Proxies for Downstream Fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1207–1239. https://doi.org/10.1145/3531146.3533180

[97] Christos Dimitrakakis, Yang Liu, David C Parkes, and Goran Radanovic. 2019. Bayesian fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 509–516.

[98] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021), 6478–6490.

[99] Jiahao Ding, Xinyue Zhang, Xiaohuan Li, Junyi Wang, Rong Yu, and Miao Pan. 2020. Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 622–629.

[100] Yasmine Djebrouni. 2022. Towards Bias Mitigation in Federated Learning. In *16th EuroSys Doctoral Workshop*.

[101] Hyungrok Do, Preston Putzel, Axel S Martin, Padhraic Smyth, and Judy Zhong. 2022. Fair Generalized Linear Models with a Convex Penalty. In *International Conference on Machine Learning*. PMLR, 5286–5308.

[102] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2796–2806.

[103] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. 2021. Fairness via representation neutralization. *Advances in Neural Information Processing Systems* 34 (2021), 12091–12103.

[104] Wei Du and Xintao Wu. 2021. Fair and Robust Classification Under Sample Selection Bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 2999–3003. https://doi.org/10.1145/3459637.3482104

[105] Flavio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2018. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing* 12, 5 (2018), 1106–1119.

[106] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[107] Jannik Dunkelau and Michael Leuschel. 2019. Fairness-Aware Machine Learning. (2019).

[108] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[109] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*. PMLR, 119–133.

[110] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[111] Michael Emmerich and André H Deutz. 2018. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural computing* 17, 3 (2018), 585–609.

[112] Simon Aagaard Enni and Ira Assent. 2018. Using Balancing Terms to Avoid Discrimination in Classification. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 947–952.

[113] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. 2021. Fairfed: Enabling group fairness in federated learning. *arXiv preprint arXiv:2110.00857* (2021).

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

[115] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. 2018. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 108–114.

[116] Elaine Fehrman, Awaz K Muhammad, Evgeny M Mirkes, Vincent Egan, and Alexander N Gorban. 2017. The five factor model of personality and evaluation of drug consumption risk. In *Data science*. Springer, 231–242.

[117] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[118] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang. 2019. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341* (2019).

[119] Benjamin Fish, Jeremy Kun, and Adám D Lelkes. 2015. Fair boosting: a case study. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Citeseer.

[120] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 144–152.

[121] Hortense Fong, Vineet Kumar, Anay Mehrotra, and Nisheeth K Vishnoi. 2021. Fairness for AUC via Feature Augmentation. *arXiv preprint arXiv:2111.12823* (2021).

[122] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 329–338.

[123] Kazuto Fukuchi, Toshihiro Kamishima, and Jun Sakuma. 2015. Prediction with model-based neutrality. *IEICE TRANSACTIONS on Information and Systems* 98, 8 (2015), 1503–1516.

[124] Kazuto Fukuchi and Jun Sakuma. 2015. Fairness-Aware Learning with Restriction of Universal Dependency using f-Divergences. *arXiv preprint arXiv:1506.07721* (2015).

[125] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R. Varshney. 2022. Causal Feature Selection for Algorithmic Fairness. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA) *(SIGMOD '22)*. Association for Computing Machinery, New York, NY, USA, 276–285. https://doi.org/10.1145/3514221.3517909

[126] Xuanqi Gao, Juan Zhai, Shiqing Ma, Chao Shen, Yufei Chen, and Qian Wang. 2022. FairNeuron: Improving Deep Neural Network Fairness with Adversary Games on Selective Neurons. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) *(ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 921–933. https://doi.org/10.1145/3510003.3510087

[127] Adriana Solange Garcia de Alford, Steven K Hayden, Nicole Wittlin, and Amy Atwood. 2020. Reducing Age Bias in Machine Learning: An Algorithmic Approach. *SMU Data Science Review* 3, 2 (2020), 11.

[128] Bhavya Ghai, Mihir Mishra, and Klaus Mueller. 2022. Cascaded Debiasing: Studying the Cumulative Effect of Multiple Fairness-Enhancing Interventions. *arXiv preprint arXiv:2202.03734* (2022).

[129] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. 2018. Online learning with an unknown fairness metric. *Advances in neural information processing systems* 31 (2018).

[130] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[131] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. 2016. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*. 2415–2423.

[132] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. 2019. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*. PMLR, 2357–2365.

[133] Przemyslaw A Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying fairness and explainability in supervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1905–1916.

[134] Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. 2021. Learning Unbiased Representations via Rényi Minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 749–764.

[135] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2021. Fairness-aware neural Rényi minimization for continuous features. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2262–2268.

[136] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2021. Fairness without the sensitive attribute via Causal Variational Autoencoder. *arXiv preprint arXiv:2109.04999* (2021).

[137] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. 2019. Fair adversarial gradient tree boosting. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1060–1065.

[138] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[139] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv preprint arXiv:1806.11212* (2018).

[140] Umang Gupta, Aaron Ferber, Bistra Dilkina, and Greg Ver Steeg. 2021. Controllable Guarantees for Fair Outcomes via Contrastive Information Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7610–7619.

[141] Philipp Hacker and Emil Wiedemann. 2017. A continuous framework for fairness. *arXiv preprint arXiv:1712.07924* (2017).

[142] Sara Hajian and Josep Domingo-Ferrer. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25, 7 (2012), 1445–1459.

[143] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[144] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[145] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. PMLR, 1939–1948.

[146] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems* 31 (2018).

[147] James M Hickey, Pietro G Di Stefano, and Vlasios Vasileiou. 2020. Fairness by Explicability and Adversarial SHAP Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 174–190.

[148] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[149] Max Hort and Federica Sarro. 2021. Did You Do Your Homework? Raising Awareness on Software Fairness and Discrimination. ASE.

[150] Max Hort, Jie Zhang, Federica Sarro, and Mark Harman. 2021. 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*.

[151] Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2022. Provably Fair Federated Learning via Bounded Group Loss. *arXiv preprint arXiv:2203.10190* (2022).

[152] Tongxin Hu, Vasileios Iosifidis, Wentong Liao, Hang Zhang, Michael Ying Yang, Eirini Ntoutsi, and Bodo Rosenhahn. 2020. Fairnn-conjoint learning of fair representations for fair decisions. In *International Conference on Discovery Science*. Springer, 581–595.

[153] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and fair classification. In *International Conference on Machine Learning*. PMLR, 2879–2890.

[154] Xiaoling Huang, Zhenghui Li, Yilun Jin, and Wenyu Zhang. 2022. Fair-AdaBoost: Extending AdaBoost method to achieve fair classification. *Expert Systems with Applications* 202 (2022), 117240.

[155] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 49–58.

[156] Alexey Ignatiev, Martin C Cooper, Mohamed Siala, Emmanuel Hebrard, and Joao Marques-Silva. 2020. Towards formal fairness in machine learning. In *International Conference on Principles and Practice of Constraint Programming*. Springer, 846–867.

[157] Eugenia Iofinova, Nikola Konstantinov, and Christoph H Lampert. 2021. Flea: Provably fair multisource learning from unreliable training data. *arXiv preprint arXiv:2106.11732* (2021).

[158] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. 2019. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 1375–1380.

[159] Vasileios Iosifidis and Eirini Ntoutsi. 2018. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke* 24 (2018).

[160] Vasileios Iosifidis and Eirini Ntoutsi. 2019. Adafair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 781–790.

[161] Vasileios Iosifidis and Eirini Ntoutsi. 2020. FABBOO-Online Fairness-Aware Learning Under Class Imbalance. In *International Conference on Discovery Science*. Springer, 159–174.

[162] Vasileios Iosifidis, Arjun Roy, and Eirini Ntoutsi. 2022. Parity-based cumulative fairness-aware boosting. *Knowledge and Information Systems* (27 Jul 2022). https://doi.org/10.1007/s10115-022-01723-3

[163] Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. 2019. Fairness-enhancing interventions in stream classification. In *International Conference on Database and Expert Systems Applications*. Springer, 261–276.

[164] Rashidul Islam, Shimei Pan, and James R Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 586–596.

[165] Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. 2020. Invariant representations through adversarial forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4272–4279.

[166] Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. 2022. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 6988–6995.

[167] Taeuk Jang, Feng Zheng, and Xiaoqian Wang. 2021. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7908–7916.

[168] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 702–712.

[169] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*. PMLR, 862–872.

[170] Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. 2022. Generalized Demographic Parity for Group Fairness. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YigKlMJwjye

[171] Jiayin Jin, Zeru Zhang, Yang Zhou, and Lingfei Wu. 2022. Input-agnostic Certified Group Fairness via Gaussian Parameter Smoothing. In *International Conference on Machine Learning*. PMLR, 10340–10361.

[172] James E Johndrow and Kristian Lum. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13, 1 (2019), 189–220.

[173] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[174] Kory D Johnson, Dean P Foster, and Robert A Stine. 2016. Impartial predictive modeling: Ensuring fairness in arbitrary models. *Statist. Sci.* (2016), 1.

[175] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2018. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 158–163.

[176] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems* 29 (2016).

[177] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2019. An algorithmic framework for fairness elicitation. *arXiv preprint arXiv:1905.10660* (2019).

[178] Sangwon Jung, Sanghyuk Chun, and Taesup Moon. 2022. Learning Fair Classifiers with Partially Annotated Group Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10348–10357.

[179] Peter Kairouz, Jiachun Liao, Chong Huang, Maunil Vyas, Monica Welfert, and Lalitha Sankar. 2022. Generating Fair Universal Representations Using Adversarial Models. *IEEE Transactions on Information Forensics and Security* 17 (2022), 1970–1985. https://doi.org/10.1109/TIFS.2022.3170265

[180] Mohammad Mahdi Kamani. 2020. Multiobjective Optimization Approaches for Bias Mitigation in Machine Learning. (2020).

[181] Mohammad Mahdi Kamani, Farzin Haddadpour, Rana Forsati, and Mehrdad Mahdavi. 2022. Efficient fair principal component analysis. *Machine Learning* (2022), 1–32.

[182] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*. IEEE, 1–6.

[183] Faisal Kamiran and Toon Calders. 2010. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Citeseer, 1–6.

[184] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[185] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874.

[186] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

[187] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences* 425 (2018), 18–33.

[188] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[189] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2013. The independence of fairness-aware classifiers. In *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 849–858.

[190] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Model-based and actual independence for fairness-aware classification. *Data Mining and Knowledge Discovery* 32, 1 (2018), 258–286.

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

[192] Kentaro Kanamori and Hiroki Arimura. 2019. 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, 3Rin211–3Rin211.

[193] Kentaro Kanamori and Hiroki Arimura. 2021. Fairness-Aware Decision Tree Editing Based on Mixed-Integer Linear Optimization. *Transactions of the Japanese Society for Artificial Intelligence* 36, 4 (2021), B–L13_1.

[194] Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. 2021. MultiFair: Multi-Group Fairness in Machine Learning. *arXiv preprint arXiv:2105.11069* (2021).

[195] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness *(Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html

[196] Thomas Kehrenberg, Zexun Chen, and Novi Quadrianto. 2019. Tuning Fairness by Marginalizing Latent Target Labels. *stat* 1050 (2019), 10.

[197] Thomas Kehrenberg, Zexun Chen, and Novi Quadrianto. 2020. Tuning fairness by balancing target labels. *Frontiers in artificial intelligence* 3 (2020), 33.

[198] Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James R Foulds. 2020. Equitable allocation of healthcare resources with fair cox models. *arXiv preprint arXiv:2010.06820* (2020).

[199] Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. 2018. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*. PMLR, 2630–2639.

[200] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 277–287.

[201] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 656–666.

[202] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. 2020. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*. PMLR, 5264–5274.

[203] Jin-Young Kim and Sung-Bae Cho. 2020. Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck.. In *SafeAI@ AAAI*. 105–112.

[204] Jin-Young Kim and Sung-Bae Cho. 2022. An information theoretic approach to reducing algorithmic bias for machine learning. *Neurocomputing* 500 (2022), 26–38. https://doi.org/10.1016/j.neucom.2021.09.081

[205] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems* 31 (2018).

[206] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 247–254.

[207] Kenji Kobayashi and Yuri Nakao. 2021. 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, 43–54.

[208] Junpei Komiyama and Hajime Shimao. 2017. Two-stage algorithm for fairness-aware machine learning. *arXiv preprint arXiv:1710.04924* (2017).

[209] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. 2018. Nonconvex optimization for regression with fairness constraints. In *International conference on machine learning*. PMLR, 2737–2746.

[210] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*. 853–862.

[211] Caitlin Kuhlman, Latifa Jackson, and Rumi Chunara. 2020. No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:2002.11836* (2020).

[212] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[213] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems* 33 (2020), 728–740.

[214] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*. IEEE, 1334–1345.

[215] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. Operationalizing Individual Fairness with Pairwise Fair Representations. *Proc. VLDB Endow.* 13, 4 (dec 2019), 506–518. https://doi.org/10.14778/3372716.3372723

[216] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. *Advances in Neural Information Processing Systems* 32 (2019).

[217] Connor Lawless, Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. 2021. Interpretable and Fair Boolean Rule Sets via Column Generation. *arXiv preprint arXiv:2111.08466* (2021).

[218] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2022), e1452.

[219] Joshua Lee, Yuheng Bu, Prasanna Sattigeri, Rameswar Panda, Gregory Wornell, Leonid Karlinsky, and Rogerio Feris. 2022. 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, 3523–3527.

[220] Joshua Lee, Yuheng Bu, Prasanna Sattigeri, Rameswar Panda, Gregory W Wornell, Leonid Karlinsky, and Rogerio Schmidt Feris. 2022. A Maximal Correlation Framework for Fair Machine Learning. *Entropy* 24, 4 (2022), 461.

[221] Joshua K Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W Wornell. 2021. Fair Selective Classification via Sufficiency. In *International Conference on Machine Learning*. PMLR, 6076–6086.

[222] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[223] Chenglu Li, Wanli Xing, and Walter Leite. 2021. 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*. 572–578.

[224] Peizhao Li and Hongfu Liu. 2022. Achieving Fairness at No Utility Cost via Data Reweighing with Influence. In *International Conference on Machine Learning*. PMLR, 12917–12930.

[225] Tianyi Li, Zhoufei Tang, Tao Lu, and Xiaoquan Michael Zhang. 2022. 'Propose and Review': Interactive Bias Mitigation for Machine Classifiers. *Available at SSRN 4139244* (2022).

[226] Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. 2023. Dark-Skin Individuals Are at More Risk on the Street: Unmasking Fairness Issues of Autonomous Driving Systems. *CoRR* abs/2308.02935 (2023).

[227] Xuran Li, Peng Wu, and Jing Su. 2022. Accurate Fairness: Improving Individual Fairness without Trading Accuracy. *arXiv preprint arXiv:2205.08704* (2022).

[228] Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, and Baowen Xu. 2022. Training Data Debugging for the Fairness of Machine Learning Software. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. 2215–2227. https://doi.org/10.1145/3510003.3510091

[229] Yueqing Liang, Canyu Chen, Tian Tian, and Kai Shu. 2022. Joint Adversarial Learning for Cross-domain Fair Classification. *arXiv preprint arXiv:2206.03656* (2022).

[230] Jixue Liu, Jiuyong Li, Lin Liu, Thuc Duy Le, Feiye Ye, and Gefei Li. 2018. FairMod-Making Predictive Models Discrimination Aware. *arXiv preprint arXiv:1811.01480* (2018).

[231] Shaofan Liu, Shiliang Sun, and Jing Zhao. 2022. Fair Transfer Learning with Factor Variational Auto-Encoder. *Neural Processing Letters* (2022), 1–13.

[232] Suyun Liu and Luis Nunes Vicente. 2022. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science* (2022), 1–25.

[233] Wenyan Liu, Xiangfeng Wang, Xingjian Lu, Junhong Cheng, Bo Jin, Xiaoling Wang, and Hongyuan Zha. 2021. Fair Differential Privacy Can Mitigate the Disparate Impact on Model Accuracy. https://openreview.net/forum?id=IqVB8e0DlUd

[234] Michael Lohaus, Michaël Perrot, and Ulrike Von Luxburg. 2020. Too relaxed to be fair. In *International Conference on Machine Learning*. PMLR, 6360–6369.

[235] Pranay Lohia. 2021. Priority-based Post-Processing Bias Mitigation for Individual and Group Fairness. *arXiv preprint arXiv:2102.00417* (2021).

[236] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias Mitigation Post-processing for Individual and Group Fairness. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2847–2851. https://doi.org/10.1109/ICASSP.2019.8682620

[237] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The Variational Fair Autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.00830

[238] Andrew Lowy, Rakesh Pavan, Sina Baharlouei, Meisam Razaviyayn, and Ahmad Beirami. 2021. FERMI: Fair Empirical Risk Minimization via Exponential R\'enyi Mutual Information. *arXiv preprint arXiv:2102.12586* (2021).

[239] Kristian Lum and James Johndrow. 2016. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077* (2016).

[240] Ling Luo, Wei Liu, Irena Koprinska, and Fang Chen. 2015. Discrimination-aware association rule mining for unbiased data analytics. In *International Conference on Big Data Analytics and Knowledge Discovery*. Springer, 108–120.

[241] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. 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*. 502–510.

[242] Ramanujam Madhavan and Mohit Wadhwa. 2020. Fairness-Aware Learning with Prejudice Free Representations. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2137–2140.

[243] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*. PMLR, 3384–3393.

[244] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*. 349–358.

[245] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems* 31 (2018).

[246] Gaurav Maheshwari and Michaël Perrot. 2022. FairGrad: Fairness Aware Gradient Descent. *arXiv preprint arXiv:2206.10923* (2022).

[247] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. 2020. There is no trade-off: enforcing fairness can improve accuracy. *arXiv preprint arXiv:2011.03173* (2020).

[248] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. 2020. Ensuring fairness beyond the training data. *Advances in neural information processing systems* 33 (2020), 18445–18456.

[249] Ricards Marcinkevics, Ece Ozkan, and Julia E Vogt. 2022. Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods. In *Machine Learning for Healthcare Conference*. PMLR.

[250] William Martin, Federica Sarro, Yue Jia, Yuanyuan Zhang, and Mark Harman. 2016. A survey of app store analysis for software engineering. *IEEE transactions on software engineering* 43, 9 (2016), 817–847.

[251] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*. PMLR, 6755–6764.

[252] Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. 2019. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*. PMLR, 4382–4391.

[253] Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. 2017. Provably fair representations. *arXiv preprint arXiv:1710.04394* (2017).

[254] Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2022. Attributing Fair Decisions with Attention Interventions. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Association for Computational Linguistics, Seattle, U.S.A., 12–25. https://doi.org/10.18653/v1/2022.trustnlp-1.2

[255] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[256] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 107–118.

[257] Alan Mishler and Edward Kennedy. 2021. FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes. *arXiv preprint arXiv:2109.00173* (2021).

[258] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2021. Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 386–400. https://doi.org/10.1145/3442188.3445902

[259] Kiarash Mohammadi, Aishwarya Sivaraman, and Golnoosh Farnadi. 2022. FETA: Fairness Enforced Verifying, Training, and Predicting Algorithms for Neural Networks. *arXiv preprint arXiv:2206.00553* (2022).

[260] Wellington Rodrigo Monteiro and Gilberto Reynoso-Meza. [n. d.]. Proposal of a Fair Voting Classifier Using Multi-Objective Optimization. ([n. d.]).

[261] Wellington Rodrigo Monteiro and Gilberto Reynoso-Meza. 2021. Proposal of a Fair Voting Classifier Using Multi-Objective Optimization.

[262] Alice Morano. 2020. *Bias mitigation for automated decision making systems*. Ph. D. Dissertation. Politecnico di Torino.

[263] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2019. Auditing and achieving intersectional fairness in classification problems. *arXiv preprint arXiv:1911.01468* (2019).

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

[265] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. 2018. Invariant representations without adversarial training. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 9102–9111.

[266] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning optimal fair policies. In *International Conference on Machine Learning*. PMLR, 4674–4682.

[267] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[268] Harikrishna Narasimhan. 2018. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1646–1654.

[269] Dang Nguyen, Sunil Gupta, Santu Rana, Alistair Shilton, and Svetha Venkatesh. 2021. Fairness improvement for black-box classifiers with Gaussian process. *Information Sciences* 576 (2021), 542–556. https://doi.org/10.1016/j.ins.2021.06.095

[270] Alejandro Noriega-Campero, Michiel A Bakker, Bernardo Garcia-Bulle, and Alex'Sandy' Pentland. 2019. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 77–83.

[271] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[272] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. 2022. Learning Fair Representation via Distributional Contrastive Disentanglement. *arXiv preprint arXiv:2206.08743* (2022).

[273] Mahbod Olfat and Anil Aswani. 2018. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1933–1942.

[274] Luca Oneto, Michele Donini, and Massimiliano Pontil. 2020. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[275] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. 2019. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 227–237.

[276] Mustafa Safa Ozdayi, Murat Kantarcioglu, and Rishabh Iyer. 2021. Bifair: Training fair models with bilevel optimization. *arXiv preprint arXiv:2106.04757* (2021).

[277] Manisha Padala and Sujit Gujar. 2020. 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*.

[278] Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. 2021. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Uncertainty in Artificial Intelligence*. PMLR, 600–609.

[279] Saerom Park, Junyoung Byun, and Joohee Lee. 2022. Privacy-Preserving Fair Learning of Support Vector Machine with Homomorphic Encryption. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3572–3583. https://doi.org/10.1145/3485447.3512252

[280] Pranita Patil and Kevin Purcell. 2022. Decorrelation-Based Deep Learning for Bias Mitigation. *Future Internet* 14, 4 (2022), 110.

[281] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM international conference on data mining*. SIAM, 581–592.

[282] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.

[283] Sikha Pentyala, Nicola Neophytou, Anderson Nascimento, Martine De Cock, and Golnoosh Farnadi. 2022. PrivFairFL: Privacy-Preserving Group Fairness in Federated Learning. *arXiv preprint arXiv:2205.11584* (2022).

[284] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. 2017. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 339–355.

[285] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 854–863.

[286] Dana Pessach and Erez Shmueli. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784* (2020).

[287] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[288] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. 2022. FAIR: Fair adversarial instance re-weighting. *Neurocomputing* 476 (2022), 14–37.

[289] Andrija Petrović, Mladen Nikolić, Miloš Jovanović, Miloš Bijanić, and Boris Delibašić. 2021. Fair classification via Monte Carlo policy gradient method. *Engineering Applications of Artificial Intelligence* 104 (2021), 104398. https://doi.org/10.1016/j.engappai.2021.104398

[290] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[291] Tao Qi, Fangzhao Wu, Chuhan Wu, Lingjuan Lyu, Tong Xu, Zhongliang Yang, Yongfeng Huang, and Xing Xie. 2022. FairVFL: A Fair Vertical Federated Learning Framework with Contrastive Adversarial Learning. *arXiv preprint arXiv:2206.03200* (2022).

[292] Shangshu Qian, Viet Hung Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, and Sameena Shah. 2021. Are my deep learning systems fair? An empirical study of fixed-seed training. *Advances in Neural Information Processing Systems* 34 (2021), 30211–30227.

[293] Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems* 30 (2017), 677–688.

[294] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2018. Neural styling for interpretable fair representations. *arXiv preprint arXiv:1810.06755* (2018).

[295] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8227–8236.

[296] Edward Raff and Jared Sylvester. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 189–198.

[297] Edward Raff, Jared Sylvester, and Steven Mills. 2018. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 243–250.

[298] Amirarsalan Rajabi and Ozlem Ozmen Garibay. 2022. Tabfairgan: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 488–501.

[299] Francesco Ranzato, Caterina Urban, and Marco Zanella. 2021. Fairness-Aware Training of Decision Trees by Abstract Interpretation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1508–1517.

[300] Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P Gummadi, and Isabel Valera. 2022. Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1421–1433.

[301] Srinivasan Ravichandran, Drona Khurana, Bharath Venkatesh, and Narayanan Unny Edakunni. 2020. FairXGBoost: Fairness-aware Classification in XGBoost. *arXiv preprint arXiv:2009.01442* (2020).

[302] Michael Redmond and Alok Baveja. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141, 3 (2002), 660–678.

[303] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. 2020. Fairness for robust log loss classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5511–5518.

[304] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9419–9427.

[305] Goce Ristanoski, Wei Liu, and James Bailey. 2013. Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1529–1532.

[306] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. 2020. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*. PMLR, 8147–8157.

[307] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. Fairbatch: Batch selection for model fairness. In *9th International Conference on Learning Representations, ICLR 2021*.

[308] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. Sample Selection for Fair and Robust Training. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

[309] Yaniv Romano, Stephen Bates, and Emmanuel Candes. 2020. Achieving equalized odds by resampling sensitive attributes. *Advances in Neural Information Processing Systems* 33 (2020), 361–371.

[310] Andrea Romei and Salvatore Ruggieri. 2011. A multidisciplinary survey on discrimination analysis.

[311] Arjun Roy, Vasileios Iosifidis, and Eirini Ntoutsi. 2022. Multi-Fair Pareto Boosting. In *International Conference on Discovery Science*. Springer.

[312] Arjun Roy and Eirini Ntoutsi. 2022. Learning to Teach Fairness-aware Deep Multi-task Learning. In *European Conference on Machine Learning and Knowledge Discovery in Databases*.

[313] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. 2020. Learning certified individually fair representations. *Advances in Neural Information Processing Systems* 33 (2020), 7584–7596.

[314] Chris Russell, M Kusner, C Loftus, and Ricardo Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in neural information processing systems*, Vol. 30. NIPS Proceedings.

[315] Bashir Sadeghi, Runyi Yu, and Vishnu Boddeti. 2019. On the global optima of kernelized adversarial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7971–7979.

[316] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. 2021. Automated feature engineering for algorithmic fairness. *Proceedings of the VLDB Endowment* 14, 9 (2021), 1694–1702.

[317] Teresa Salazar, Miriam Seoane Santos, Helder Araújo, and Pedro Henriques Abreu. 2021. FAWOS: Fairness-Aware Oversampling Algorithm Based on Distributions of Sensitive Attributes. *IEEE Access* (2021).

[318] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*. 793–810.

[319] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair pca: One extra dimension. In *Advances in Neural Information Processing Systems*. 10976–10987.

[320] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. 2020. Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision*. Springer, 746–761.

[321] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. 2020. Intra-processing methods for debiasing neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 2798–2810.

[322] Nicolas Schreuder and Evgenii Chzhen. 2021. Classification with abstention but without disparities. In *Uncertainty in Artificial Intelligence*. PMLR, 1227–1236.

[323] Marco Scutari, Francesca Panero, and Manuel Proissl. 2021. Achieving Fairness with a Simple Ridge Penalty. *arXiv preprint arXiv:2105.13817* (2021).

[324] Emel Seker, John R Talburt, and Melody L Greer. 2022. Preprocessing to Address Bias in Healthcare Data. *Studies in Health Technology and Informatics* 294 (2022), 327–331.

[325] Shubham Sharma, Alan H. Gee, David Paydarfar, and Joydeep Ghosh. 2021. FaiR-N: Fair and Robust Neural Networks for Structured Data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 946–955. https://doi.org/10.1145/3461702.3462559

[326] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 358–364.

[327] Martin B. Short and George O. Mohler. 2022. A fully Bayesian tracking algorithm for mitigating disparate prediction misclassification. *International Journal of Forecasting* (2022). https://doi.org/10.1016/j.ijforecast.2022.05.008

[328] Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. 2022. Fair Representation Learning through Implicit Path Alignment. *arXiv preprint arXiv:2205.13316* (2022).

[329] Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2022. GetFair: Generalized Fairness Tuning of Classification Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 289–299.

[330] Arashdeep Singh, Jashandeep Singh, Ariba Khan, and Amar Gupta. 2022. Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair. *Machine Learning and Knowledge Extraction* 4, 1 (2022), 240–253.

[331] Agnieszka Słowik and Léon Bottou. 2021. Algorithmic Bias and Data Bias: Understanding the Relation between Distributionally Robust Optimization and Data Curation. *arXiv preprint arXiv:2106.09467* (2021).

[332] P. Snel and S. van Otterloo. 2022. Practical bias correction in neural networks: a credit default prediction case study. *Computers and Society Research Journal* 3 (2022).

[333] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2164–2173.

[334] Ezekiel Soremekun, Mike Papadakis, Maxime Cordy, and Yves Le Traon. 2022. Software Fairness: An Analysis and Survey. *arXiv preprint arXiv:2205.08809* (2022).

[335] Haipei Sun, Kun Wu, Ting Wang, and Wendy Hui Wang. 2022. Towards Fair and Robust Classification. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 356–376.

[336] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[337] Supreme Court of the United States. 2009. *Ricci v. DeStefanoo*. Vol. 557.

[338] Vinith M Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. 2022. When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction. *arXiv preprint arXiv:2206.02058* (2022).

[339] Jared Sylvester and Edward Raff. 2020. Trimming the Thorns of AI Fairness Research. *IEEE Data Eng. Bull.* 43, 4 (2020), 74–84.

[340] Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. 2020. Learning fair representations for kernel models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 155–166.

[341] Guanhong Tao, Weisong Sun, Tingxu Han, Chunrong Fang, and Xiangyu Zhang. 2022. 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.*

[342] Maryam Tavakol. 2020. Fair Classification with Counterfactual Learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2073–2076.

[343] Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. 2022. Fairness-Aware Configuration of Machine Learning Libraries. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) *(ICSE ’22)*. Association for Computing Machinery, 909–920. https://doi.org/10.1145/3510003.3510202

[344] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*. PMLR, 6373–6382.

[345] Ana Valdivia, Javier Sánchez-Monedero, and Jorge Casillas. 2021. How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems* 36, 4 (2021), 1619–1643.

[346] Benjamin van Giffen, Dennis Herhausen, and Tobias Fahse. 2022. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research* 144 (2022), 93–106.

[347] Sahil Verma, Michael Ernst, and Rene Just. 2021. Removing biased data to improve fairness and accuracy. *arXiv preprint arXiv:2102.03054* (2021).

[348] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[349] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[350] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2022. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Trans. Knowl. Discov. Data* (jul 2022). https://doi.org/10.1145/3551390 Just Accepted.

[351] Guanchu Wang, Mengnan Du, Ninghao Liu, Na Zou, and Xia Hu. 2022. Mitigating Algorithmic Bias with Limited Annotations. *arXiv preprint arXiv:2207.10018* (2022).

[352] Hao Wang, Berk Ustun, and Flavio Calmon. 2019. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*. PMLR, 6618–6627.

[353] Hao Wang, Berk Ustun, Flavio P Calmon, and SEAS Harvard. 2018. Avoiding disparate impact with counterfactual distributions. In *NeurIPS Workshop on Ethical, Social and Governance Issues in AI.*

[354] Jingbo Wang, Yannan Li, and Chao Wang. 2022. Synthesizing Fair Decision Trees via Iterative Constraint Solving. In *International Conference on Computer Aided Verification*. Springer, 364–385.

[355] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 526–536.

[356] Jialu Wang, Xin Eric Wang, and Yang Liu. 2022. Understanding Instance-Level Impact of Fairness Constraints. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 23114–23130. https://proceedings.mlr.press/v162/wang22ac.html

[357] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. 2020. Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems* 33 (2020),

5190–5203.

[358] Xiaoqian Wang and Heng Huang. 2019. Approaching machine learning fairness through adversarial network. *arXiv preprint arXiv:1909.03013* (2019).

[359] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. 2021. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1748–1757.

[360] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. 2020. Optimized score transformation for fair classification. *Proceedings of Machine Learning Research* 108 (2020).

[361] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: a trade-off revisited. In *NeurIPS*.

[362] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).

[363] Claes Wohlin. 2014. 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*. 1–10.

[364] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. In *Conference on Learning Theory*. PMLR, 1920–1953.

[365] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. Semi-FairVAE: Semi-supervised Fair Representation Learning with Adversarial Variational Autoencoder. *arXiv preprint arXiv:2204.00536* (2022).

[366] Songhua Wu, Mingming Gong, Bo Han, Yang Liu, and Tongliang Liu. 2022. Fair classification with instance-dependent label noise. In *Conference on Causal Learning and Reasoning*. PMLR, 927–943.

[367] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. Fairness-aware classification: Criterion, convexity, and bounds. *arXiv preprint arXiv:1809.04737* (2018).

[368] Ziwei Wu and Jingrui He. 2022. Fairness-Aware Model-Agnostic Positive and Unlabeled Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1698–1708. https://doi.org/10.1145/3531146.3533225

[369] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. *Advances in neural information processing systems* 30 (2017).

[370] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

[371] Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*. 594–599.

[372] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 570–575.

[373] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 1401–1406.

[374] Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair class balancing: enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1715–1724.

[375] Jenny Yang, Andrew AS Soltan, Yang Yang, and David A Clifton. 2022. Algorithmic Fairness and Bias Mitigation for Clinical Machine Learning: Insights from Rapid COVID-19 Diagnosis by Adversarial Learning. *medRxiv* (2022).

[376] Mehdi Yazdani-Jahromi, AmirArsalan Rajabi, Aida Tayebi, and Ozlem Ozmen Garibay. 2022. Distraction is All You Need for Fairness. *arXiv preprint arXiv:2203.07593* (2022).

[377] I-Cheng Yeh and Che-hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications* 36, 2 (2009), 2473–2480.

[378] Xiaoxin Yin and Jiawei Han. 2003. CPAR: Classification based on predictive association rules. In *Proceedings of the 2003 SIAM international conference on data mining*. SIAM, 331–335.

[379] Zhe Yu. 2021. Fair Balance: Mitigating Machine Learning Bias Against Multiple Protected Attributes With Data Balancing. *CoRR* abs/2107.08310 (2021). arXiv:2107.08310 https://arxiv.org/abs/2107.08310

[380] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1gdkxHFDH

[381] Mikhail Yurochkin and Yuekai Sun. 2021. SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness. In *International Conference on Learning Representations*.

[382] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[383] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

[384] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-Based Notions of Fairness in Classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 228–238.

[385] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.

[386] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2020. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery* 34, 1 (2020), 163–200.

[387] Vladimiro Zelaya, Paolo Missier, and Dennis Prangle. 2019. Parametrised data sampling for fairness optimisation. *KDD XAI* (2019).

[388] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[389] Xianli Zeng, Edgar Dobriban, and Guang Cheng. 2022. Bayes-Optimal Classifiers under Group Fairness. *arXiv preprint arXiv:2202.09724* (2022).

[390] Xianli Zeng, Edgar Dobriban, and Guang Cheng. 2022. Fair Bayes-Optimal Classifiers Under Predictive Parity. *arXiv preprint arXiv:2205.07182* (2022).

[391] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 335–340.

[392] Hantian Zhang, Xu Chu, Abolfazl Asudeh, and Shamkant B Navathe. 2021. OmniFair: A Declarative System for Model-Agnostic Group Fairness in Machine Learning. In *Proceedings of the 2021 International Conference on Management of Data*. 2076–2088.

[393] Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 3675–3685.

[394] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[395] Jiang Zhang, Ivan Beschastnikh, Sergey Mechtaev, and Abhik Roychoudhury. 2022. Fair Decision Making via Automated Repair of Decision Trees. In *International Workshop on Equitable Data and Technology (FairWare '22)*.

[396] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* (2020).

[397] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 3929–3935.

[398] Lu Zhang, Yongkai Wu, and Xintao Wu. 2018. Achieving Non-Discrimination in Prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm, Sweden) *(IJCAI'18)*. AAAI Press, 3097–3103.

[399] Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C Weiss, and Wolfgang Nejdl. 2021. 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*. 245–256.

[400] Wenbin Zhang and Eirini Ntoutsi. 2019. FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 1480–1486.

[401] Wenbin Zhang, Xuejiao Tang, and Jianwu Wang. 2019. On fairness-aware learning for non-discriminative decision-making. In *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1072–1079.

[402] Wenbin Zhang and Jeremy C Weiss. 2021. Fair decision-making under uncertainty. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 886–895.

[403] Wenbin Zhang and Jeremy C Weiss. 2022. Longitudinal fairness with censorship. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12235–12243.

[404] Wenbin Zhang, Jeremy C Weiss, Shuigeng Zhou, and Toby Walsh. 2022. Fairness amidst non-iid graph data: A literature review. *arXiv preprint arXiv:2202.07170* (2022).

[405] Xueru Zhang and Mingyan Liu. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 525–555.

[406] Yue Zhang and Arti Ramesh. 2020. Learning Fairness-aware Relational Structures. *arXiv preprint arXiv:2002.09471* (2020).

[407] Chen Zhao, Feng Chen, and Bhavani Thuraisingham. 2021. Fairness-Aware Online Meta-learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2294–2304.

[408] Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, and Feng Chen. 2022. Adaptive Fairness-Aware Online Meta-Learning for Changing Environments. *arXiv preprint arXiv:2205.11264* (2022).

[409] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. 2020. Conditional Learning of Fair Representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Hkekl0NFPr

[410] Han Zhao and Geoff Gordon. 2019. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems* 32 (2019).

[411] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. *EMNLP* (2018), 4847–4853.

[412] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2021. You Can Still Achieve Fairness Without Sensitive Attributes: Exploring Biases in Non-Sensitive Features. *arXiv preprint arXiv:2104.14537* (2021).

[413] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. 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*. 1433–1442.

[414] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. 2021. Learning Bias-Invariant Representation by Cross-Sample Mutual Information Minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15002–15012.

[415] Indre Žliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).

[416] Indre Žliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling conditional discrimination. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 992–1001.