In-Processing Modeling Techniques for Machine Learning Fairness: A Survey

MINGYANG WAN, Department of Computer Science and Engineering, Texas A&M University, USA
DAOCHEN ZHA, Department of Computer Science, Rice University, USA
NINGHAO LIU, Department of Computer Science, University of Georgia, USA
NA ZOU, Department of Engineering Technology and Industrial Distribution, Texas A&M University, USA

Machine learning models are becoming pervasive in high-stakes applications. Despite their clear benefits in terms of performance, the models could show discrimination against minority groups and result in fairness issues in a decision-making process, leading to severe negative impacts on the individuals and the society. In recent years, various techniques have been developed to mitigate the unfairness for machine learning models. Among them, in-processing methods have drawn increasing attention from the community, where fairness is directly taken into consideration during model design to induce intrinsically fair models and fundamentally mitigate fairness issues in outputs and representations. In this survey, we review the current progress of in-processing fairness mitigation techniques. Based on where the fairness is achieved in the model, we categorize them into explicit and implicit methods, where the former directly incorporates fairness metrics in training objectives, and the latter focuses on refining latent representation learning. Finally, we conclude the survey with a discussion of the research challenges in this community to motivate future exploration.

CCS Concepts: • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**.

Additional Key Words and Phrases: Machine Learning Fairness, Bias Mitigation, Disparate Impact, Disparate Treatment

1 INTRODUCTION

Machine learning models are becoming pervasive in real-world applications and have been increasingly deployed in high-stakes decision-making processes, such as loan management [62], job applications [69], and criminal justice [9]. Despite the clear benefits brought by machine learning techniques, e.g., strong prediction ability by automatically discovering effective patterns in data, the resultant models could show discrimination against certain groups of people and lead to fairness issues in practice. For example, it is reported that the deployed automated risk assessment tools for criminal offenders have significant racial disparities: black defendants are almost twice more likely to be falsely flagged as future criminals than white defendants [5]. The problem is partly caused by the bias that already exists in datasets and could be further amplified by models [82], leading to severe negative impacts on the individuals and the society. To mitigate unfairness in machine learning models, researchers have developed various techniques at different stages of model development [46, 47, 50, 59, 60, 63, 64, 68], including pre-processing, in-processing and post-processing methods. Pre-processing [13, 46, 59, 60, 63] and post-processing [50, 64, 68] are straightforward strategies to mitigate unfairness: pre-processing methods focus on adjusting the training

Authors' addresses: Mingyang Wan, w1996@tamu.edu, Department of Computer Science and Engineering, Texas A&M University, USA; Daochen Zha, daochen.zha@rice.edu, Department of Computer Science, Rice University, USA; Ninghao Liu, ninghao.liu@uga.edu, Department of Computer Science, University of Georgia, USA; Na Zou, nzou1@tamu.edu, Department of Engineering Technology and Industrial Distribution, Texas A&M University, USA.

1556-4681/2022/7-ART $15.00
https://doi.org/10.1145/3551390

data distribution to balance the sensitive groups, while post-processing methods calibrate the prediction results after model training.

Unlike the above techniques, in-processing methods directly incorporate fairness into model design, which can induce intrinsically fair models and fundamentally mitigate fairness issues in machine learning models [30, 31, 40, 48, 76]. First, in-processing techniques can address the problem of bias amplification in model training: the tendency that the model exacerbates biases in the training data. This problem is often caused by the algorithm and cannot be solely attributed to the data [34, 81]. In-processing methods directly take fairness into consideration in model optimization so that the converged model could achieve fairness even with biased data as input [18]. Second, in-processing methods can effectively fine-tune the representations from pre-trained models to mitigate the bias without huge re-training efforts. While many pre-trained deep learning models [25, 41, 42] have demonstrated their performance power, they inevitably encode or amplify bias, leading to potential unfair consequences in downstream applications. It is with huge efforts and sometimes even infeasible to re-train the models for unfairness mitigation. Instead, in-processing methods can be used to fine-tune the representations in pre-trained models. For example, a contrastive learning framework was proposed to debias the representations without re-training for BERT [19], an indispensable pre-trained component in the modern natural language processing models [25]. Third, it remains a challenge to develop effective in-processing solutions. The existing algorithms mainly focus on addressing the fairness issue under a single and known sensitive attribute. There exist application scenarios where many of the existing methods will fail. For instance, it is crucial to address fairness issues when sensitive attributes information is not disclosed or available.

Complementing the previous survey papers that mainly present high-level overview of machine learning fairness and general taxonomies of pre-processing, in-processing, and post-processing methods [18, 29, 61], we aim to summarize and categorize the key ideas behind the existing in-processing methods for motivating future exploration of algorithmic fairness. In this article, we survey through the in-processing methods for machine learning fairness and categorize them into explicit and implicit mitigation methods based on where the fairness is achieved in the model. Explicit mitigation is achieved by formulating the objective function with fairness constraints or regularizers, which are often inspired by the fairness measurements and designed to enforce the predictions to be less dependent on the sensitive attributes. The explicit methods are often flexible and easy to implement as they only require minimum modifications to objective functions. Some example regularizers include co-variance relationships [47], absolute correlation regularization [11], Wasserstein-1 distances [43], etc. Alternatively, implicit mitigation focuses on debiasing the representations so that the resulting predictions do not show discrimination towards a minority group or individuals. The implicit approaches often target deep learning models, where learning representations is crucial. For example, they can be used to fine-tune the representations in pre-trained models to debias downstream applications. Some strategies that fall into this category include adversarial learning [31], contrastive learning [19], disentangled representation learning [58], etc.

The remainder of this article is structured as follows. Section 2 gives an introduction of the fairness problem and some representative measurements to quantify the (un)fairness. Section 3 and Section 4 summarize the existing research on explicit and implicit mitigation of unfairness, respectively. Section 5 provides general discussions on the differences of the existing in-processing methods. Section 6 discusses the research challenges in this community. Finally, we conclude the article in Section 7.

2 FAIRNESS IN MACHINE LEARNING

In this section, we introduce the fairness problem in machine learning, measurements for quantifying the degree of bias for different types of fairness, and an overview of the taxonomy for in-processing techniques.

Table 1. Main symbols and definitions.

Symbol	Definition
f_θ	A machine learning model that maps attributes to predictions with parameters θ.
$\mathbf{x} \in \mathbb{R}^d$	A vector of attributes with a dimension of d.
$x_s \in \mathbb{R}$	The sensitive attribute.
$\mathbf{s}$	A tuple representing all the values of (multiple) sensitive attributes.
$\hat{y} \in \{-1, 1\}$	A binary prediction that indicates negative and positive outcomes for -1 and 1, respectively.
$y \in \{-1, 1\}$	The ground-truth label.
$\mathcal{D}$	The training dataset.
$L(\mathcal{D}; \theta)$	The objective function of the machine learning model.
$\mathbf{z} \in \mathbb{R}^l$	The encoded latent representation with a dimension of l.
$R(\mathcal{D}; \theta)$	Fairness regularizer.
$\Omega(\mathcal{D}; \theta)$	Fairness constraint.
$\|\theta\|_2^2$	L_2 regularizer.
$[\cdot]_+$	The operator that ignores the parts less than zero.
$d(\cdot, \cdot)$	Distance between two samples in attribute space.
$D(\cdot, \cdot)$	Distance between two samples in prediction space.
$\mathbf{x}_+$	A positive sample in contrastive learning
$\mathbf{x}_-$	A negative sample in contrastive learning
$\mathrm{MI}(\cdot, \cdot)$	Mutual information.
$\mathbf{x}_{\bar{s}}$	Non-sensitive attributes.
$\mathbf{b}$	The sensitive latent factors.
$\hat{x}_s$	The predicted sensitive attribute.
A_ϕ	An adversary network with parameters ϕ.
$L(\mathcal{D}; \phi)$	The adversarial loss for $\hat{x}_s$ and x_s.
$L_r(\mathcal{D}; \theta)$	The reconstruction loss.

2.1 Fairness Problem

We first discuss fairness in general, and distinguish the concepts of fairness and bias from the psychometrics view. Then, we describe different strategies to improve fairness and narrow the scope to in-processing methods. Finally, we formally define the fairness problem from the computational perspective, followed by a supervised binary classification task as an example to describe the underlying fairness issues.

2.1.1 *Definition of Fairness in Machine Learning.*

Fairness is a social and subjective concept whose definition varies across different cultures and societies. In the context of machine learning, unfairness often refers to the situation where a model shows discrimination against certain groups of people (e.g., races and genders). Another term that frequently appears is *bias*, which is often used interchangeably with fairness in the machine learning literature [61]. However, these two concepts are different from the psychometrics view. A recent study has provided a clear and comprehensive discussion to distinguish the two [14]. Unlike fairness, bias refers to any systematic error that affects the performance of different groups in different ways. Bias is defined as either contamination of the measurement of a construct (e.g., by adding unnecessary gender information) or a deficiency in the measurement (e.g., by not capturing all aspects of the construct which are relevant). A biased model could be either fair or unfair. For example, in a hypothetical experiment performed in [67], a college admission decision algorithm that is given knowledge

about demographics could help admit more people from underrepresented races. The algorithm is inherently biased (while fair) because it uses demographic information. We refer interested readers to [14] for a more comprehensive discussion.

In this article, *fairness* of a machine learning model is defined in the context of a specific measurement. A model cannot be fair under all fairness measurements, since those fairness definitions are mutually exclusive [12]. Explicit methods will often result in fair models because they directly incorporate fairness metrics into learning objectives. Some implicit methods, such as contrastive learning [19] and disentanglement [58], only seek to reduce the reliance on sensitive attributes in representations, which result in a debiased model. The fairness performance depends on the downstream tasks and the fairness measurements (detailed in Section 2.2) to be used. Domain experts and application areas help determine which fairness metric to choose.

2.1.2 *Different Views on Fairness in Machine Learning.*

Fairness in the context of machine learning is very complex and can be interpreted from different views. Consider the machine learning process as a pipeline consisting of data collection, feature engineering, model training, model prediction, decision making, etc. Three dimensions of fairness have been discussed in the literature [14, 73], including distributive, interactional, and procedural fairness. Distributive fairness focuses on the outcomes brought by machine learning (e.g., jobs or resource allocation). Interactional fairness regards how people perceive the explanations, rationales, and justifications of the decisions. Procedural fairness refers to the fairness of each element of the decision-making, i.e, each step of the machine learning pipeline. In this article, we focus on the distributive perspective, which is related to in-processing mitigation methods.

From the technical perspective, algorithms to improve fairness can be categorized into pre-processing, in-processing, and post-processing methods according to where the intervention is applied. Pre-processing regards adjusting the data before feeding them into the model, while post-processing is applied after modeling, such as calibrating the prediction results. In-processing is applied to the design and training of models, which can induce intrinsically fair models. In this article, we focus on in-processing methods, which can fundamentally mitigate fairness issues even with biased inputs by taking fairness into consideration in model optimization. While increasing research efforts have been dedicated to the in-processing methods in recent years, it remains a challenge to develop effective solutions for many scenarios (e.g., the lack of sensitive attribute information). This article aims to summarize the ideas of the existing in-processing methods to motivate future exploration.

2.1.3 *Fairness from the Computational Perspective.*

Without loss of generality, we consider a binary classification problem. Formally, the task is to learn a mapping $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}$, which takes as input the attributes $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ (including a sensitive attribute $x_s \in \mathcal{X}_s \subseteq \mathbb{R}$) for each individual, where θ is the model weights and d denotes the number of attributes. Typically, we only consider a single sensitive attribute. It is also worth noting that researchers have studied fairness problems where the sensitive attribute information is not fully available [40], or there are multiple sensitive attributes [48]. In the following discussions, we focus on the most common scenario with a single known sensitive attribute. Given all the attributes, the model will make a binary prediction $\hat{y} \in \{-1, 1\}$ that indicates negative and positive outcomes for -1 and 1, respectively (we use $y \in \{-1, 1\}$ to denote the ground truth). The model is then often trained on a training dataset $\mathcal{D}$ using an objective function $L(\mathcal{D}; \theta)$ with the goal that the trained model could make predictions for unlabeled individuals in a hold-out test set. For a deep learning model, the mapping f_θ usually consists of an encoder that maps $\mathbf{x}$ into a latent representation $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^l$, where l is the dimension of the latent space, and a decoder that maps $\mathbf{z}$ into $\hat{y}$. We summarize the symbols used throughout this article in Table 1. Regardless of the adopted models, the models could be unfair in the use of the sensitive attribute x_s. Next, we instantiate the above problem definition with the *Adult* dataset, which will serve as an example in our discussion to facilitate understanding.

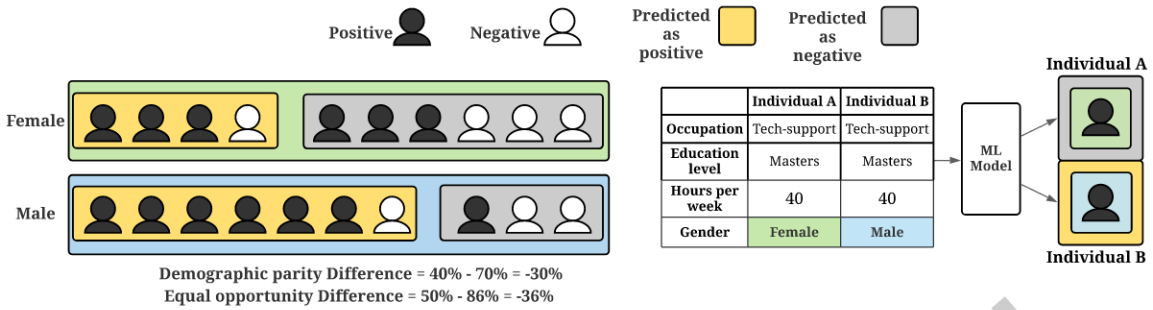

Fig. 1. Disparate impact (left) and disparate treatment (right) on the *Adult* dataset, where a binary classifier is trained to predict high salary (positive) or low salary (negative). Disparate impact addresses group-level bias, where the two sensitive groups (male and female) are treated differently according to a fairness metric (e.g., true positive rates are significantly different across groups). Disparate treatment focuses on individual-level bias, where two individuals with different sensitive attributes (i.e., gender) but similar non-sensitive attributes of occupation, education level and hours work per week are treated differently.

The task of the *Adult* dataset is to predict whether a person's annual salary is higher (positive) or lower (negative) than 50K dollars. The attributes in *Adult* are tabular, describing the characteristics of each individual. They consist of many non-sensitive attributes, such as occupation, education level, and hours per week, and some sensitive attributes, such as gender and race. We consider a single sensitive attribute x_s, which is binary and $x_s \in \{male, female\}$. From the computational perspective, the fairness problem can be generally grouped into two categories: *disparate impact* and *disparate treatment*, which approach the fairness problem from the group- and individual-level, respectively [84].

- **Disparate Impact:** It refers to the situation where the model disproportionately discriminates certain groups [84], even if the model does not explicitly leverage the sensitive attribute to make predictions but rather on some proxy attributes. The left-hand side of Figure 1 provides an example of disparate impact in the *Adult* dataset. According to the prediction results, the model is unfair because it tends to predict male instances as positive with a higher probability (i.e., 0.7) than females as positive (i.e., 0.4). In the example, the true-positive rate is 50% in the female group while the rate is 86% in the male group. There are various ways and perspectives to measure disparate impact, which will be introduced in details later.
- **Disparate Treatment:** Unlike disparate impact that focuses on group-level discrimination, disparate treatment refers to unfairness at the level of individuals. The underlying intuition is that a model should not treat individuals with similar attributes differently. The right-hand side of Figure 1 gives an example of disparate treatment in the *Adult* dataset. The two individuals A and B have very similar background information such as the same occupation, equal education level, and identical work hours per week, while gender is the only different attribute between them. In this example, the model is unfair because the prediction for individual A is negative while it is positive for individual B. Such a phenomenon usually suggests that the model could have undesirably leveraged sensitive attributes, such as gender information in this example, to make predictions.

2.2 Fairness Measurements

As many fairness measurements have been proposed, choosing the fairness measurement to be applied is crucial for detecting and mitigating model unfairness. The chosen measurement decides how fairness is defined, and it reflects the expectation of target applications. This subsection focuses on some representative ones that quantify

Table 2. A summary of fairness measurements.

Category	Mechanism	Sensitive Attribute	Measurement	Description
Group	Parity	Known	Demographic parity [23]	Equal percentages of positive outcome
	Confusion Matrix	Known	Equal opportunity [39]	Equal true positive rate
			Equalized odds [9]	Equal true positive rate and false positive rate
			Overall accuracy equality [9]	Equal accuracy
			Treatment equality [9]	Equal false negative rate / false positive rate
			Equalizing disincentives [45]	Equal true positive rate - false positive rate
	Worst-off utility	Unknown	Rawlsian Max-Min [70]	The lowest the utility is maximized
	Cluster balance	Known	Fair cluster [20]	The ratios of the groups are balanced for each cluster
Individual	Lipschitz property	Known	Fairness through awareness [30]	Similar individuals have similar outcomes
	Causal mode	Known	Counterfactual fairness [54]	Same prediction for actual/counterfactual individuals
Hybrid	Bounding	Known	Differential fairness [34]	Positive/negative outcomes bounded across sub-groups

the disparate impact (group fairness measurements) or the disparate treatment (individual fairness measurements), respectively. Additionally, we introduce a hybrid criteria that measures the fairness of subsets of multiple sensitive attributes. Table 2 summarizes the measurements according to group/individual measurements, the mechanisms, and whether sensitive attribute is known. We still use *Adult* dataset as the example for illustration.

Note that, due to the subjectivity of fairness, different fairness measurements could be suitable for different scenarios. We provide more discussions in Section 5. In this article, we assume that the appropriate fairness measurements are already chosen in downstream applications, and focus on surveying the design of unfairness mitigation techniques.

2.2.1 *Group Fairness Measurements.*

Group fairness measures the difference of model predictions on two or more groups. The group that an individual belongs to is indicated by its sensitive attributes. We use s_i and s_j to denote two different sensitive groups associated with a sensitive attribute. For the example task of salary prediction, s_i and s_j represent female and male groups, respectively. In some scenarios, there could be more than two sensitive groups. In this case, we can often aggregate the measurement values for each pair of groups as an overall measurement. Following the above notations, we summarize some representative group fairness measurements as below, where the first measurement is parity-based (i.e., considering positive rates), the next five measurements are based on confusion matrix (i.e., considering aspects of true positive rates, true negative rates, false positive rates, and false negative rates), the next measurement is based on worst-off utility (i.e., the worst value of a metric, such as accuracy, AUC, across groups), and the final one is defined for clustering problems.

- **Demographic parity [23]**: The percentages of a positive outcome across different sensitive groups (denoted by x_s) should be the same, i.e., $p(\hat{y} = 1|x_s = s_i) = p(\hat{y} = 1|x_s = s_j)$. In the example of Figure 1, 40% of females are predicted as positive, while the ratio is 70% for male. Due to the large gap between the two ratios, it is probable that the model violates demographic parity.
- **Equal opportunity [39]**: Different groups should have equal true positive rates, i.e., $p(\hat{y} = 1|x_s = s_i, y = 1) = p(\hat{y} = 1|x_s = s_j, y = 1)$. In Figure 1, three out of six truly positive individuals in the female group are predicted as positive, with a true positive rate of 50%; in contrast, the true positive rate of the male group is 86%, which is significantly higher than 50%. Therefore, it is probable that the model fails to guarantee equal opportunity.
- **Equalized odds [9]**: Different groups should have equal true positive and false positive rates, i.e., $p(\hat{y} = 1|x_s = s_i, y = 1) = p(\hat{y} = 1|x_s = s_j, y = 1)$ and $p(\hat{y} = 1|x_s = s_i, y = -1) = p(\hat{y} = 1|x_s = s_j, y = -1)$. This measurement is more restrictive than demographic parity and equalized odds since we require both true

and false positive rates to be the same. It is often used when we strongly care about predicting the positive outcomes correctly.

- **Overall accuracy equality [9]:** The accuracies across the sensitive groups are the same, i.e., $p(\hat{y}=1|x_s=s_i, y=1)+p(\hat{y}=-1|x_s=s_i, y=-1)=p(\hat{y}=1|x_s=s_j, y=1)+p(\hat{y}=-1|x_s=s_j, y=-1)$. Unlike the above measurements, overall accuracy equality emphasizes both true positive and true negative rates. This can be used in the scenarios where true negatives are as desirable as true positives.
- **Treatment equality [9]:** The ratios of false negative prediction to false positive prediction should be equal across groups, i.e., $\frac{p(\hat{y}=1|y=-1,x_s=s_i))}{p(\hat{y}=-1|y=1,x_s=s_i))}=\frac{p(\hat{y}=1|y=-1,x_s=s_j))}{p(\hat{y}=-1|y=1,x_s=s_j))}$. This measurement can be used when incorrectly classifying an individual results in a bigger loss.
- **Equalizing disincentives [45]**: The difference between true positive rates and false positive rates should be equal across groups, i.e., $p(\hat{y}=1|y=1,x_s=s_i))-p(\hat{y}=1|y=-1,x_s=s_i))=p(\hat{y}=1|y=1,x_s=s_j))-p(\hat{y}=1|y=-1,x_s=s_j))$. This measurement has strong emphasis on classifying positives both correctly and incorrectly.
- **Rawlsian Max-Min fairness principle [70]:** Let $U(\theta, s)$ be the utility of the sensitive group s with weights θ. The Rawlsian Max-Min fairness can be quantified as $\arg\max_\theta \min_s U(\theta, s)$. This measurement encourages maximizing the utility of the group with the lowest utility, where the utility metric can be accuracy, AUC, etc. While the definition of this measurement does not explicitly consider sensitive groups, it implicitly treats the individuals with the lowest utility as the minority group. It is commonly used in the scenarios where the sensitive information is unknown [40, 55].
- **Fair clustering [20]:** It is defined for quantifying unfairness in clustering problems. Let $N(\hat{y}=c|x_s=s_i)$ denote the number of individuals that are assigned to cluster c with sensitive attribute s_i. The balance of the cluster c is defined as the minimum ratios of the numbers of individuals in each group, i.e., $\text{balance}(c)=\min\{\frac{N(\hat{y}=c|x_s=s_i)}{N(\hat{y}=c|x_s=s_j)}, \frac{N(\hat{y}=c|x_s=s_j)}{N(\hat{y}=c|x_s=s_i)}\}$. The generated clusters are considered fair if for every cluster c, $\text{balance}(c)\geq\delta$, where δ is a threshold.

2.2.2 *Individual Fairness Measurements.*

Instead of measuring fairness across different sensitive groups, individual fairness measurements consider the fairness for each individual with the intuition that similar individuals should be treated as similarly as possible. Some representative scenarios include:

- **Fairness through awareness [30]:** Any two individuals who have similar non-sensitive attributes should receive a similar outcome. Let $d(\mathbf{x}_a, \mathbf{x}_b)$ define the difference between the attributes of two individuals $\mathbf{x}_a$ and $\mathbf{x}_b$. If $d(\mathbf{x}_a, \mathbf{x}_b)$ is small, then $D(\hat{y}_a, \hat{y}_b)$ should also be small, where $D(\cdot,\cdot)$ computes the prediction difference. Formally, the goal is to bound $D(\hat{y}_a, \hat{y}_b)$ and make the model satisfy the $(D, d)-Lipschitz$ property: $D(\hat{y}_a, \hat{y}_b)\leq d(\mathbf{x}_a, \mathbf{x}_b)$. The specific formulation of $D(\cdot,\cdot)$ and $d(\cdot,\cdot)$ are usually determined by the task at hand.
- **Counterfactual fairness [54]:** It is defined with causal models. Let (U, V, F) be a causal model, where, U denotes the set of background variables, V denotes the set of observable variables, and F denotes the set of functions associated with V. $\hat{y}$ is counterfactually fair towards an individual if the prediction is the same for the actual individual and a counterfactual individual that belongs to a different sensitive group, i.e., for any context $\mathbf{x}$, $p(\hat{y}_{x_s\leftarrow s_i}(U)=y|\mathbf{x})=p(\hat{y}_{x_s\leftarrow s_j}(U)=y|\mathbf{x})$.

2.2.3 *Hybrid Measurements.*

Some hybrid measurements, instead of measuring the difference in the group- or the individual-level, focus on the subsets of multiple sensitive attributes. We use $\mathbf{s}_i$ and $\mathbf{s}_j$ to denote two different tuples representing all the sensitive attribute values. For example, if we consider both gender and race in the *Adult* dataset, an instantiation of such tuple can be $\{male, black\}$ or $\{female, white\}$. We introduce a representative measurement as below.

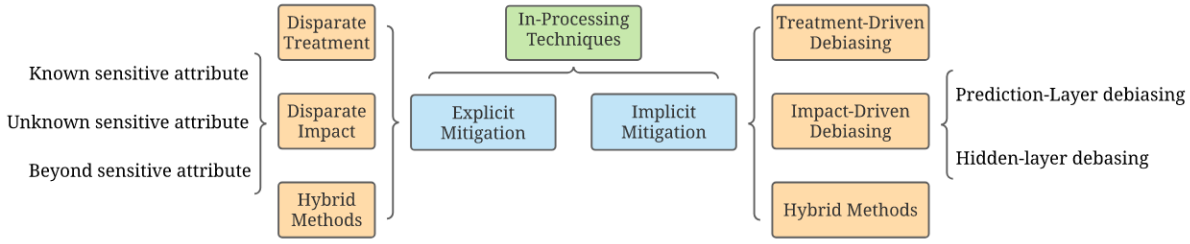

Fig. 2. Structure of the survey.

- **Differential fairness [34]:** A model f_θ is ϵ-differentially fair if $e^{-\epsilon} \leq \log \frac{p(\hat{y}=0|\mathbf{s}_i)}{p(\hat{y}=0|\mathbf{s}_j)} \leq e^{\epsilon}$ and $e^{-\epsilon} \leq \log \frac{p(\hat{y}=1|\mathbf{s}_i)}{p(\hat{y}=1|\mathbf{s}_j)} \leq e^{\epsilon}$ for all $\mathbf{s}_i$ and $\mathbf{s}_j$. This criteria states that the predictions are similar and bounded within the range $(e^{-\epsilon}, e^{\epsilon})$ regardless of the combinations of the sensitive attributes.

2.3 Survey Structure

This survey will give an overview of the recent in-processing techniques developed for various machine learning tasks. We will summarize the core ideas and design principles regardless of the specific applications. We divide existing techniques into two categories:

- **Explicit methods:** Research in this line often focuses on explicitly revising the training objective $L(\mathcal{D}; \theta)$. One common strategy is to add fairness related constraints and solve the problem with constrained optimization techniques. Another strategy is adding a regularization term into the objective function to penalize the unfairness behaviors. We provide detailed categorizations in Section 3.
- **Implicit methods:** An alternative direction is to debias the latent representation $\mathbf{z}$ to implicitly remove bias from the model. The intuition is that if $\mathbf{z}$ is less biased, the resulting predictions $\hat{y}$ from $\mathbf{z}$ will also be less biased. We provide detailed categorizations in Section 4.

We further categorize explicit and implicit methods into several sub-categories based on whether the algorithms are designed to mitigate the disparate impact (i.e., group fairness/bias), disparate treatment (i.e., individual fairness/bias), or both of them. Under each sub-category, based on the existing techniques, we either provide a high-level summary of the key ideas (e.g., a unified formula) with some representative instances, or further divide them based on the scenarios that the algorithms can tackle. An overview of the taxonomy used in this article is summarized in Figure 2.

Complementing the previous survey papers that mainly present a high-level overview of machine learning fairness and general taxonomies [18, 29, 61], we provide a more fine-grained and technical categorization by focusing on in-processing methods. Our categorization provides a clearer understanding of where the fairness is achieved in existing techniques (in training objectives or representations, in which layers of the representations), what fairness goals they can achieve (e,g., disparate impact or disparate treatment), and what scenarios they can tackle (e.g., whether sensitive attribute is known).

2.4 Strategy for Paper Collection and Category Determination

We identified the relevant papers and determined the categories based on the following iterative procedure. Firstly, we identified a small set of papers by using the query “fair machine learning” along with some related keywords, such as “bias”, “discrimination”, and "unfairness", to narrow down the search results. Note that we did not explicitly put the keyword “in-processing” in the query since many in-processing papers do not explicitly mention this term in the title or abstract. Instead, we manually examined them to collect the relevant ones.

Among the papers returned from the search, we mainly considered the fairness work from the computational perspective but not other disciplines such as law or social science. We focused on those important papers which are either published at major machine learning or data mining conferences (such as ICML, NeurIPS, ICLR, KDD, WSDM, CIKM, WWW, SIGIR, ACL, EMNLP, NAACL, AAAI, IJCAI, etc.) or journals (such as JMLR, TPAMI, TKDD, TKDE, etc.) in the past five years (as most of the papers are published very recently without many citations), or have more than 50 citations (as some pioneering works are very popular even though they are not published in the above venues). In addition to directly searching for papers from the venues, we also traced the related work section in recent papers to identify additional papers. We only considered the technical part of in-processing methods in the paper and do not include the discussions of legal issues of fairness. Secondly, we categorize the papers according to our taxonomy (e.g., explicit and implicit methods).

We iteratively repeated the above two steps to collect more papers following the categories determined in the previous iterations, and gradually refined the categories and developed sub-categories based on the newly collected papers. Finally, the identified taxonomy converged to the one shown in Figure 2. Note that rather than desperately trying to cover all the work in this domain, we only selected the most representative papers under each sub-category to make our discussion focused. In essence, some sub-categories (e.g., adversarial debiasing) can have many extensions and variants. In this case, we manually traced the citation graph to identify the most influential papers for discussion. We have identified more than 100 papers studying in-processing methods for unfairness mitigation. Among them, we removed the extensions or pure applications of basic approaches (e.g., adversarial debiasing) where the basic approach was also present in the list, resulting in 22 explicit (Table 3) and 16 implicit (Table 4) representative techniques.

3 EXPLICIT UNFAIRNESS MITIGATION

Explicit unfairness mitigation can be achieved by explicitly adding fairness regularizers or constraints to the objective functions of machine learning models. The high-level ideas are illustrated in Figure 3. In this section, we summarize different explicit mitigation methods according to whether they tackle disparate impact or disparate treatment. Additionally, researchers have also studied hybrid methods that address both of them. A summary of the surveyed explicit mitigation methods is provided in Table 3.

3.1 Disparate Impact Mitigation

The existing explicit mitigation methods for addressing disparate impact can be categorized based on whether sensitive attribute information is known. Beyond these, some other studies have also explored fairness problems that are irrelevant to sensitive attributes.

3.1.1 Fairness with known sensitive attribute information.

A large body of existing work has investigated how to mitigate disparate impact when the sensitive information, such as the race or gender information for each individual, is available in the training data. A typical strategy is to guide the model with fairness regularization so that predictions are less dependent on the sensitive attributes. A general objective function with fairness learning is formulated as:

$$L(\mathcal{D};\theta) + \lambda\|\theta\|_2^2 + \eta R(\mathcal{D};\theta), \tag{1}$$

where the first term is the loss of the main task (e.g., classification or regression), the second term is a standard regularizer on model parameters, the third term is a designed fairness regularizer, and η and λ are hyperparameters. Here, the fairness regularizer $\mathrm{R}(\mathcal{D};\theta)$ is the key to achieve the model fairness.

The core idea of the regularizer design is to quantify and minimize the correlation between the sensitive attributes and the prediction. A representative strategy to quantify such correlation is based on mutual information. [47] introduced *prejudice index* (PI in short) to quantify the degree of dependence between a sensitive variable

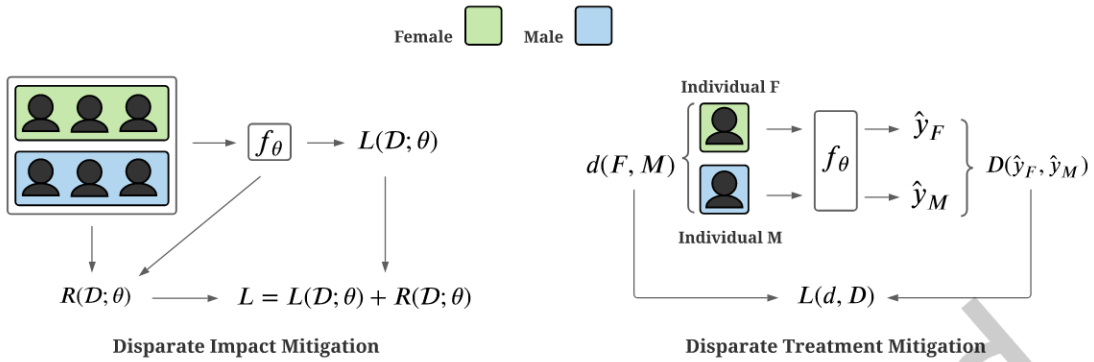

Fig. 3. An illustration of explicit mitigation of bias. The left-hand side shows an example of disparate impact mitigation with regularization terms. In addition to the original loss $L(\mathcal{D};\theta)$, a fairness term $R(\mathcal{D};\theta)$ is formulated to enforce group fairness. The existing studies mainly differ in how $R(\mathcal{D};\theta)$ is designed. The right-hand side illustrates disparate treatment mitigation. Here, $d(\cdot,\cdot)$ and $D(\cdot,\cdot)$ measure the distance between any two individuals in attribute space and prediction space, respectively. The objective is to minimize the discrepancy between these two distances for individuals with similar non-sensitive attributes but different sensitive attributes. In this example, a female (F) and a male (M) with similar non-sensitive attributes are enforced to have similar predictions.

and a target variable. It is defined as the sample distributions over a given sample set of the mutual distribution between the sensitive groups and the predictions. By minimizing the mutual information between these two, we enforce the predictions less dependent on the sensitive attributes. In this sense, the model tends to be fair across groups. Although this paper focuses on simple linear regression and naive Bayesian classifiers, the idea is general and could be also applied to other models. Other notable regularizers include minimizing Wasserstein-1 distances between the classifier outputs and sensitive information [43], absolute correlation regularization that takes a simplified view by minimizing the absolute value of the correlation between the sensitive groups and the predictions [11], and residual between the clicked and unclicked item and the members in groups in recommender systems [10], etc.

Another idea of designing objective functions is to minimize the loss under fairness constraints, which can be formulated as a constraint optimization problem:

$$\begin{aligned} \min_{\theta} \quad & L(\mathcal{D};\theta) + \lambda\|\theta\|_2^2 \\ \text{s.t.} \quad & \Omega(\mathcal{D};\theta) < 0, \end{aligned} \tag{2}$$

where $L(\mathcal{D};\theta)$ is the loss of the main task, the second term is a standard regularizer, and Ω is a fairness constraint that is usually defined as a convex function. An example is to formulate Ω as the covariance between the sensitive attributes and the signed distance from the feature vectors to the decision boundary [85, 86]. In this example, Ω is a convex function because the signed distance is convex with respect to θ. The convex nature ensures that Ω will not increase the complexity of the training.

Complementing the above strategies that mainly focus on model predictions, some other studies alternatively investigate the scenarios where ground truth is available in the historical data. A noteworthy example is mitigating *disparate mistreatment* [84], which aims to achieve similar misclassification rates for different sensitive groups. Disparate mistreatment is usefull in many real-world scenarios, such as assigning risk scores to criminal offenders. For example, ProPublica[1] collects data of criminal offenders in Broward County, Florida, during 2013-2014. The

[1]https://github.com/propublica/compas-analysis

Table 3. A summary of explicit mitigation methods. *Regularization* refers to adding a penalty term to the objective function to make the optimal solution fairer. *Constraint optimization* refers to optimizing the parameters with the original objective in the presence of some (hard) constraints on the parameters. Note that regularization is also regarded as a soft constraint in the literature (i.e., a constraint which is preferred but not required to be satisfied). In this survey, we distinguish these two by grouping the papers with hard constraint into constraint optimization and papers with soft constraint into regularization.

Fairness Goal	Training Scheme	Sensitive Attribute	Method
Disparate impact	Regularization	Known	Prejudice index [47], Absolute correlation [11] Wasserstein fair [43], Pairwise comparisons [10] Fair regression [3], Fair decision trees [4]
	Constraint optimization	Known	Flexible mechanism [86], Tractable constraints [85], Convex-concave [76]
		Unknown	DRO [40], ARL [55], Proxy Fairness [38]
		Beyond	Paper matching [52]
Disparate treatment	Regularization	Known	Controlled direct effect [26], Fair decision trees [4], Convex fair regression [8]
	Constraint optimization	Known	Fairness Through Awareness [30], Logit pairing [36] Counterfactual fairness [54]
Hybrid methods	Constraint optimization	Known	subgroup fairness [48], rich subgroup fairness[48] Maxmin-Fair Ranking [35]

data consists of offenders' demographic features, such as gender, race, and age, as well as the offenders' criminal history. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool[2] is then used to assign a risk score to each offender. The data also provides ground truth on whether the offenders have recidivated within two years. An individual is misclassified if a high (low) risk score is assigned to an offender who has not (has) recidivated within two years. The automated decision making system may suffer from disparate mistreatment if the misclassification rates for different sensitive groups (e.g., race) differ. To address this, a constraint optimization problem is formulated to bound the difference of the misclassification rates between two sensitive groups. The problem can then be efficiently solved with convex-concave programming [76]. In this way, the classifier learns the optimal decision boundary subject to the fairness constraints, which facilitates fair treatments across sensitive groups. However, such constraints only bound the overall misclassification rates, while specific individuals within a group can still be treated unfairly.

Recent studies have explored fairness in unsupervised learning tasks, where the label information is unavailable. The existing studies in this line of research mainly focus on clustering, which is often considered as the most common unsupervised learning problem. One representative work, Fair Clustering [20], defines fairness in clustering under the disparate impact doctrine. Specifically, it defines the balance of a cluster as the ratio between the numbers of individuals belonging to each sensitive group within the cluster. A clustering algorithm is considered fair if all the generated clusters are relatively balanced. Unfortunately, achieving the optimal solution under such fairness definition is NP-hard because the number of combinations grows exponentially with more individuals. This work develops an approximation algorithm based on *fairlets*, which refers to small and fair clusters. The key idea is to first obtain some farilets, which can be solved in polynomial time by relating it to the graph covering problem. Then it clusters the fairlets using standard clustering algorithms. Since the fairlets are balanced, the obtained clusters (each cluster is formed by multiple farilets) will also be balanced. A follow-up

[2]https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

work extends Fair Clustering by making the first step (i.e., fairlets decomposition) more efficient by mapping the data points into a tree metric. We refer interested readers to [6] for more details.

3.1.2 Fairness without sensitive attribute information.
All the above studies assume that sensitive attributes, such as gender and race, are available in the dataset. However, collecting sensitive attributes itself is sometimes infeasible due to privacy issues. To deal with these situations, researchers have investigated how to mitigate unfairness when we do not have full access to sensitive attributes. The main idea of these studies is that, while we do not have direct access to the unobserved sensitive attributes, they are often correlated with some other observed attributes. For example, the races could be correlated with the zip codes [24]. As such, we could mitigate the racial unfairness even without knowing the races by treating the individuals with a certain zip code as the minority group. This is because if zip codes can divide the individuals in a similar way as races, mitigating unfairness with zip codes will have similar effects as races. Note that whether this idea is effective highly depends on the degree of the correlations between the sensitive attribute and the other attributes. Some attributes could have strong correlations with the sensitive attribute while the others may only have weak correlations. For example, it is well known that vocal pitch correlates with gender. In this case, the vocal pitch could serve as an almost perfect substitute for gender. In contrast, in the case of races and zip codes, while they have a good correlation in general, the exact correlation may differ in different cities and countries. Thus, a specific zip code may not necessarily correspond to exactly one race. The degree of correlations will determine the fairness performance of the mitigation algorithms since they highly rely on such correlations to achieve fairness.

Proxy Fairness [38] explores this idea by addressing fairness on proxy groups, i.e., the groups identified based on the correlated observed attributed, as substitutes for the sensitive groups. In the above example, we can define the proxy groups based on the zip code as substitutes for races. In this way, any existing fairness mitigation techniques could be applied to the proxy groups, which could also indirectly address the fairness issues in the true sensitive groups. In [38], they formulate a constraint optimization problem on the proxy groups and surprisingly observe that such a simple strategy could work well in practice. However, the authors also note that the effectiveness of the idea depends on the choice of fairness metric and the alignment between the proxy groups and the true sensitive groups. This approach requires careful selection of the proxy features and could only meet some specific needs.

Another direction focuses on improving Rawlsian Max-Min Fairness [70], which aims at maximizing the utility of the worst-case group, i.e., the group with the lowest utility. The methods in this line of research also rely on the correlated observed attributes for unfairness mitigation. The main idea is to identify the regions with higher classification errors using the observed attributes and give a higher weight to the individuals within these regions to optimize the worst-case performance. One noteworthy training strategy is based on *Distributionally Robust Optimization* (DRO) [40]. The assumption behind is that the performance of minority groups is sacrificed when achieving the overall performance since the contribution of minority groups to the overall loss is significantly less than that of majority groups, that is, the minority group is underrepresented during the training. To address this issue, DRO minimizes

$$\mathbb{E}_P[L(\mathcal{D};\theta) - \eta]_+^2, \tag{3}$$

where $L(\mathcal{D};\theta)$ is the classification loss, $[\cdot]_+$ drops the instances with loss smaller than η, and η is a dual variable controlling the threshold for small loss. This objective function can optimize the worst-case performance in that both the dual variable η and the squared term can up-weight the samples with high losses. However, DRO may suffer from the risk of focusing on optimizing the performance of outliers since the losses of outliers are usually above the threshold.

A follow-up work proposes to address this limitation with adversarial reweighted learning [55]. In this work, the Rawlsian Max-Min Fairness objective is formulated as a minimax problem of a zero-sum game between

two players, where one player tries to minimize the reweighted loss function, and the adversary player aims to maximize the loss by assigning the weights. The weight assignment is achieved by a neural network that takes as input the observed features and outputs a value between 0 to 1 as the weight. During the training process, the adversary tends to assign higher weights to the instances with higher losses, and the other player will focus more on these instances to optimize the worst-case performance. A linear adversary is used and shown to deliver the best performance since it can mitigate the impact of the noisy outliers.

3.1.3 Beyond sensitive attributes.
While previous research efforts mainly center on sensitive attributes, there exist fairness scenarios that are irrelevant to sensitive attributes and instead focus on the fairness issues in specific application domains.We note that such fairness problems can be very specific may not be encountered in other domains.

Paper matching problem [52] is a typical scenario that falls into this subcategory. It aims at automatically matching reviewers to the papers in the reviewing process. Instead of addressing fairness problems across sensitive groups, they focus on fairness regarding papers and reviewers. Specifically, they identify two fairness goals in paper reviewing: (1) the reviewers assigned to a paper should collectively possess sufficient expertise, and (2) each reviewer should have a reasonable workload in terms of the number of the assigned papers. On one hand, each paper is treated as an individual, and the assignment should be in fair in terms of the expertise of the assigned reviewers. On the other hand, each reviewer is expected to be fairly treated in terms of workloads. Both of these fairness goals do not involve sensitive information . The paper matching problem can be formulated as a global optimization problem that maximizes the sum of affinities, which are scores indicating the affinity of reviewer-paper pairs. Then, the two fairness goals are achieved via adding constraints to the optimization process.

3.2 Disparate Treatment Mitigation

Disparate treatment mitigation can be achieved by penalizing the discrimination towards individuals with different sensitive attributes in objectives, i.e., similar individuals in different groups should be treated as similarly as possible. The methods in this research line often rely on task-specific metrics to measure the similarity between individuals. The idea is applicable to both classical machine learning models and deep models, where fairness is usually achieved by formulating a constrained optimization problem or adding regularization terms to the loss functions.

Fairness Through Awareness [30] is a representative approach for traditional classifiers. The key idea is to formulate a fair classifier as a constrained optimization problem, which minimizes the (linear) classification loss subject to a fairness constraint. Specifically, given a distance metric between individuals $d(\cdot,\cdot)$ and a distance measure between outputs $D(\cdot,\cdot)$, for any pair of individuals $\mathbf{x}_a$ and $\mathbf{x}_b$, the model is optimized subject to $(D,d)-$ *Lipschitz* property:

$$D(f_\theta(\mathbf{x}_a), f_\theta(\mathbf{x}_b)) \leq d(\mathbf{x}_a, \mathbf{x}_b), \tag{4}$$

where f_θ is the model and $f_\theta(\mathbf{x})$ denotes the prediction result of $\mathbf{x}$. The above constraint expresses that the distance of the outputs is bounded by the distance of the corresponding inputs. In this way, similar individuals (i.e., the distance of inputs is small) will also have similar outcomes (i.e., the distance of outputs is also small since it is bounded by the distance of inputs), which will lead to individual fairness. If f_θ is linear, the optimization problem can then be expressed as a linear programming that can be solved efficiently. Here, the two distance metrics $D(\cdot,\cdot)$ and $d(\cdot,\cdot)$ are key to achieve the goal of "treating similar individuals similarly". Since $f_\theta(\mathbf{x})$ is essentially a probability distribution, a straightforward choice for $D(\cdot,\cdot)$ is a statistical distance metric between

two probability distributions P and Q on a finite domain A (i.e., there are A possible outcomes), denoted by

$$D_{\text{tv}}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|, \tag{5}$$

where $D_{\text{tv}}(P, Q)$ is the total variation distance, and two very different distributions tend to have larger statistical distances and vice versa. $d(\cdot, \cdot)$ is a task-specific metric describing the similarity level of two individuals, which highly depends on the task.

An alternative idea for disparate treatment mitigation is *counterfactual fairness* [54], which leverages the causal framework to model the sensitive attributes. A decision is considered fair to an individual if the prediction is the same for the actual and counterfactual worlds. One representative example of counterfactual fairness modeling is counterfactual logit pairing [36], which is a robustness term that penalizes the norm of the logit difference between a pair of an actual example and their counterfactual example. Another similar work proposes to remove the direct effect of the sensitive attribute with regularization [26], which aims to minimize the difference between an actual individual and a counterfactual individual that belongs to a different sensitive group.

Other studies have designed constraints or regularizers for different kinds of tasks or models with linear or non-linear mappings. Some representative examples include adding fairness regularizers for achieving fairness in regression tasks [8], and learning fair decision trees [4].

3.3 Hybrid Methods for Explicit Bias Mitigation

Disparate impact mainly addresses group-level fairness issues but could not handle individual-level ones. Similarly, disparate treatment focuses on individual-level rather than group-level fairness. However, there are hybrid scenarios that fall beyond these two categories. This subsection reviews two representative ones: subgroup fairness that defines fairness in the subgroup-level and hybrid methods that attempt to achieve both group- and individual-level fairness.

Rather than defining fairness purely at the group- or individual-level, subgroup fairness [48] defines fairness over a combinatorially large number of subgroups. [48] gives a toy example as a failure of group-level fairness, called fairness gerrymandering. Consider a problem with two sensitive attributes including race and gender, where men/women and whites/blacks are equally represented in the target population. Suppose a binary classifier predicts positive if and only if an individual is a black man or a white woman, then the positive rate is 50% within each sensitive group. That is, the classifier will be considered to be fair under a disparate impact metric; however, it is unfair to black men or white women. Motivated by this example, [48] proposes to address subgroup fairness, where the classifier is expected to satisfy the fairness constraints for each possible subgroup. The proposed solution formulates the problem as a two-player zero-sum game between a Learner and an Auditor, where the Learner's decision space is the classifier itself, and the Auditor has the decision space of subgroups. This game can be solved with Fictitious Play with provably asymptotic convergence. Their follow-up work [49] further empirically shows that this approach works well in various datasets and demonstrates that subgroup fairness can be satisfied with the reasonable cost in terms of computational resources and accuracy. Unlike group-level or individual-level fairness, this research line focuses on fairness at the subgroup-level. Intuitively, subgroup fairness will reduce to group-level fairness if there is only one subgroup and will become individual-level fairness if there is only one individual for each subgroup.

One idea to achieve both disparate impact and disparate treatment is based on constrained optimization. A notable example is *Maxmin-Fair Ranking* [35], which aims to optimize individual fairness under group fairness constraints. Specifically, this paper focuses on a ranking problem, where people or items are assigned scores that indicate the relevance to the query. It is important that the ranking system should be fair since the result will have a direct tangible impact on the people or items being ranked. This paper formulates a minimax problem to minimize individual unfairness while enforcing the group fairness constraints that are derived from sensitive

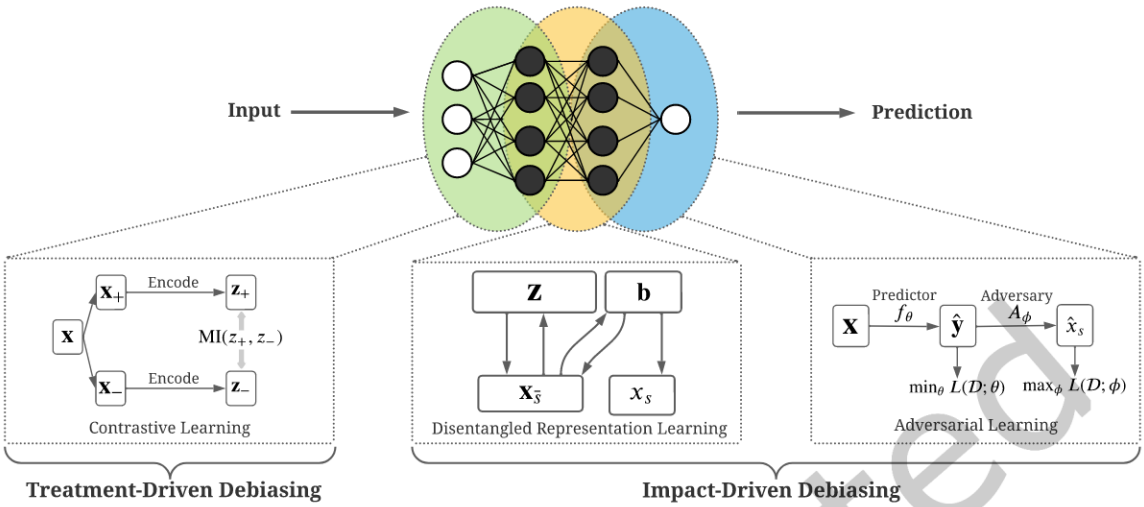

Fig. 4. Some representative techniques for implicit mitigation of bias, where different techniques often focus on different phases of the model. Contrastive learning methods mitigate disparate treatment by minimizing the mutual information between the representations obtained from positive samples ($\mathbf{x}_+$) and negative samples ($\mathbf{x}_-$), which have similar non-sensitive attributes but different sensitive attributes (it often operates on the input and the hidden layers). Disentangled learning approaches mitigate the disparate impact by learning disentangled sensitive and non-sensitive representations and using the obtained non-sensitive representations for downstream applications (it often only operates on the hidden layers). Adversarial learning strategies use an auxiliary adversary network to minimize the predictive probability of the sensitive attributes from the representations and mitigate the bias extracted in the representation (it often operates on the hidden layers and the output layers).

attributes:

$$r' \in \arg\max_{r \in \mathcal{S}} \min_{\mathbf{x} \in X} V(r, \mathbf{x}), \tag{6}$$

where r or r' denotes a ranking of the individuals, $\mathcal{S}$ denotes the subset of the rankings that satisfy the group fairness constraints, $V(r, \mathbf{x})$ denotes the utility of $\mathbf{x}$ if we place $\mathbf{x}$ according to rank r; that is, our goal is to maximize the worst-case utility across individuals under the group fairness constraints. Based on this formulation, the paper further improves the individual treatment with randomization, i.e., deriving a probability distribution over valid rankings instead of obtaining a deterministic solution. The optimization can be solved with combinatorial optimization in polynomial time.

4 IMPLICIT UNFAIRNESS MITIGATION

Implicit unfairness mitigation refers to the algorithms which detect and eliminate discrimination via representations. These approaches are mainly designed for deep models, where learning representations is crucial. They seek to remove bias from the learned representations such that the resulting predictions will also be less biased. Note that, as discussed in Section 2, an unbiased prediction may not necessarily be fair. Nevertheless, in most cases, it has been reported in many implicit mitigation methods that removing bias aligns with the common fairness goals [19, 37, 87]. Figure 4 illustrates some representative techniques for implicit mitigation, which can be grouped into impact-driven methods and treatment-driven methods. The impact-driven methods focus on debiasing the model to reduce group-level unfairness, while the treatment-driven approaches focus

Table 4. A summary of implicit mitigation methods.

Fairness Goal	**Training Techniques**	**Method**
Impact-Driven	Adversarial Learning (Prediction-layer)	Adversarial Debiasing [87] Adversarial Recidivism Application [80]
	Adversarial Learning (Hidden-layer)	Censored Representation [31] Transferable Representations [60] Re-embeds word vector [78], Fair Word Embedding [32]
	Disentanglement (Hidden-layer)	Flexibly Fair [37] Disentangled Representations [58]
Treatment-Driven	Contrastive Learning	Contrastive Debiasing [19], Multi-CLRec [89] Conditional Contrastive [79], Fair Graph [53] Fairness-aware Data Augmentation [53]
Hybrid methods	Disentanglement	Flexibly Fair [37], Intersected Fair [65]
	Adversarial Learning	Fair Graph Embeddings [15]

on the unfairness at the individual level. Some representative techniques for impact-driven methods include adversarial learning and disentangled representation learning, where the former uses an adversary to remove the sensitive information from the representations, and the latter decorrelates the sensitive and non-sensitive information to preserve unbiased representations for downstream applications. Alternatively, contrastive learning achieves treatment-driven debiasing by enforcing similar representations for positive and negative sample pairs. In addition, some hybrid methods have been proposed to tackle both disparate impact and disparate treatment. A summary of the implicit mitigation methods is tabulated in Table 4.

4.1 Impact-Driven Debiasing

The goal of impact-driven debiasing is to obtain representations that cannot infer sensitive attributes. As such, the resultant predictions will be less biased towards any sensitive groups. Based on which parts of the representations are targeted, the existing studies mainly fall into prediction-layer debiasing and hidden-layer debiasing.

4.1.1 Prediction-Layer debiasing.

Prediction-layer debiasing aims at mitigating the bias at the prediction layer. This idea is often achieved in a "backward" way through adversarial learning. The optimization problem is formulated as a minimax game between a predictor and an adversary, where the goal is to maximize the predictor's ability to predict the labels based on the representations and minimize the adversary's ability to predict the sensitive attributes from the representations. In this sense, accurate predictions are achieved without relying on sensitive attribute information.

Consider a predictor f_θ that is trained to accomplish the task of predicting $\hat{y}$ given $\mathbf{x}$ with the loss function $L(\mathcal{D};\theta)$ using a gradient-based method (illustrated in the right part of Figure 4). The output layer of the predictor then serves as the input of another adversary network A_ϕ that aims to predict sensitive attributes x_s. The adversary is jointly trained with the predictor to ensure that the predictor satisfies the desired fairness condition but still performs well on the prediction task. More formally, the objective can be written as the following minimax problem:

$$\min_{\theta} \max_{\phi} L(\mathcal{D};\theta) + L(\mathcal{D};\phi), \tag{7}$$

where $L(\mathcal{D};\phi)$ is the adversarial loss that indicates the prediction error of sensitive attributes.

A representative work [87] has investigated this idea and presented a general formulation of using adversarial learning to mitigate the unwanted bias. This paper shows that the framework can be used to achieve various fairness conditions by adjusting the inputs of the adversarial network based on the fairness definitions. To achieve demographic parity, the adversary will predict the sensitive attributes from the prediction $\hat{y}$ since demographic parity only cares about the percentages of a positive outcome. The adversary can get both $\hat{y}$ and the ground truth y for equalized odds because it relies on the ground truth to calculate true positive and false positive rates. For equal opportunity, we can only feed the training data with $y = 1$ to the adversary so that only the true positive rates are considered. This work demonstrated the effectiveness of such adversarial learning framework in mitigating the racial unfairness on income prediction tasks. Note that, although the adversarial network designs are tailored for the fairness measurements, we still use the term *debaising* since the core principle is still enforcing the representations to be less biased.

A follow-up work further verified the generality of the adversarial learning framework by applying it to criminal history datasets [80]. The task is to predict the offenders' recidivism based on the attributes of the offenders, which is widely adopted for a decision-making system in the USA Criminal Justice. It is found that there are severe racial biases in criminal history datasets. They use an adversarial network to predict race from the recidivism. The adversarial training penalizes the prediction network if the race is predictable from the recidivism prediction. By adding this adversary network, we can counteract racial discrimination in the criminal history dataset. It is reported that, compared with the standard recidivism prediction model, the adversarial training can decrease the false positive gap from 0.05 to 0.01, and false negative gap from 0.27 to 0.02, with no clear drop of accuracy [80].

4.1.2 *Hidden-Layer debiasing with adversarial learning.*

Hidden-layer debiasing aims to mitigate the bias on latent representations learned in intermediate layers. A representative work is learning *censored representation* [31]. The model consists of four components: an encoder, a decoder, an adversary, and a predictor. The encoder maps input $\mathbf{x}$ to a latent representation $\mathbf{z}$ while the decoder takes as input $\mathbf{z}$ and the sensitive attributes to reconstruct $\mathbf{x}$. The adversary tries to predict the sensitive attributes from latent representation $\mathbf{z}$. Finally, the predictor predicts the output based on $\mathbf{z}$. The objective can be formulated as

$$\min_{\theta} \max_{\phi} L(\mathcal{D}; \theta) + L(\mathcal{D}; \phi) + L_{\mathrm{r}}(\mathcal{D}; \theta), \tag{8}$$

where $L(\mathcal{D}; \phi)$ denotes the adversarial loss, and $L_{\mathrm{r}}(\mathcal{D}; \theta)$ denotes the reconstruction loss. Without $L_{\mathrm{r}}(\mathcal{D}; \theta)$, the objective is very similar to the adversarial learning used in prediction-layer debiasing in Eq. 7 with the only difference that the input of the adversary is a hidden-layer. The parameters are optimized with alternate gradient descent and gradient ascent steps. In the first step, the adversary is fixed, and we update the weights of the encoder, the decoder, and the predictor with gradient descent. Then, the adversary takes a step to maximize the loss with the other components fixed. In this way, the representation converges to the point that removes sensitive information.

A follow-up work explores learning adversarially fair and transferable representations [60]. They assume a model with similar components, including an encoder, a decoder, an adversary, and a predictor network. Unlike [31], they address the specific choices of adversarial objectives so that the representations can more closely align with the debiasing goals. In addition to classification tasks, they also demonstrate that the method can remove the bias in transfer learning. It has been shown that this method has theoretical grounding for certain debiasing objectives and is effective in classification and transfer learning tasks.

The idea of adversarial debiasing has been extended and explored on unsupervised tasks. A notable example is Deep Fair Clustering (DFC) [56]. The key idea is to separately conduct the clustering algorithm on each sensitive group to obtain pseudo assignments of the instances. These pseudo assignments can be treated as the predicted

"classes". In this way, we can apply adversarial debiasing designed for supervised learning to the clustering problem. Their empirical results suggest this simple adaptation of adversarial debiasing to clustering problems can make the generated clusters significantly more balanced.

Recently, hidden-layer debiasing methods have also been developed for sentiment analysis in natural language processing. Demographic identity terms, i.e., the words that can reflect the demographical information of an individual, can show unfair sentiment polarity. For example, the national origin terms like American, Mexican, or names that tend to belong to African American demographics like Darnell should be neutral with respect to sentiment. However, these words are shown to have positive or negative sentiment in the embedding space. To address this problem, in [78], the authors remove the discriminated sentiment correlations from the word embedding through adversarial training, which re-embeds word vectors without distorting the meaning. Intuitively, this regime minimizes the ability of the adversary to predict the sentiment polarity while maximizing the learner's ability to preserve the word vector after debiasing. By combining these two objectives in gradient update, they obtain the debiased word embeddings. Another work has explored similar methods for fair word embedding [32]. This work focuses on reaching text-driven representations that are blind to the attributes we wish to protect. They show that adversarial training is effective for debiasing but may introduce instability in training, which makes the adversary score untrusted. Potential directions to improve the results include tuning the capacity and weights of the adversary and using several adversaries as a combination.

4.1.3 Hidden-Layer debiasing with disentanglement.

Disentangled representation learning is another strategy for hidden-layer debiasing. The main assumption is that the sensitive and non-sensitive attributes are entangled in the generated representation by an underlying mixing mechanism. To make the prediction results fair across sensitive groups, disentangled representation learning aims at extracting the sensitive information out of the representation and preserving the effective, informative and unbiased representation for downstream applications. Disentangled representation learning often considers a generative model that takes the form [66]:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) \prod_i p(z_i), \tag{9}$$

where it is assumed that the observation $\mathbf{x}$ is controlled by l independent factors $z_1, z_2, ..., z_l$. The goal of disentangled representation learning is to learn the $\mathbf{z}$ such that each dimension in $\mathbf{z}$ corresponds to no more than one semantic factor that controls the variation of the data. For example, in image analysis, one dimension of $\mathbf{z}$ could extract the eyeglasses of a human face, and the other dimensions represent other parts of the human face [77]. Generally, disentangled representation learning has shown advantages in various applications in terms of its interpretability and generalizability of models [7].

In the context of model fairness, the existing work has studied disentangled representation learning in two settings with or without sensitive attributes. On one hand, disentangled representation learning is an effective way to achieve fairness with known sensitive attributes [37]. Consider the encoder and decoder structure in the middle part of Figure 4, and let $\mathbf{b}$ denote the sensitive latent factors. The model training objective is to learn the encoder and decoder while encouraging low MI($\mathbf{b}$, $\mathbf{z}$), where MI$(\cdot, \cdot)$ denotes mutual information. The model is rewarded for decorrelating the latent dimensions of $\mathbf{b}$ and $\mathbf{z}$ from each other. The model can be trained in an end-to-end fashion by combining a prediction loss. One benefit of this approach is that it is very flexible; we can consider multiple sensitive attributes in training and debias with different subsets of the sensitive attributes at the test time.

On the other hand, the effectiveness of disentangled representation learning has also been demonstrated even when the sensitive attributes are unobserved in the training data [58]. It applies standard disentangled representation techniques to the model without considering sensitive attributes and conducts a large number of experiments

to study the relationship between the disentanglement scores and fairness. Specifically, the standard disentangled representation learning will encourage each factor (i.e., an element in the representation vector) of the learned representation to be semantically independent of the other factors. Here, the specific semantic depends on the context of the task. For example, for the task of learning disentangled face representations [77], disentanglement aims to enable each element of the representation vector to control a different facial attribute, such as the skin color, hair length, sunglasses color, etc. The disentangled representations have many desirable properties, such as interpretability (i.e., we can explain each facial attribute based on the disentangled representations) [2], and better generalization ability (the disentangled representations could capture domain invariant features and be less susceptible to overfitting) [17]. The experiments of [58] show that the disentangled representations obtained by the standard disentangled representations learning procedure can improve the fairness in the downstream task. They surprisingly found that disentanglement is consistently correlated with increased fairness. A potential reason for why it works is that disentanglement makes the non-sensitive semantic factors less dependent of the sensitive attributes, which inherently makes the predictions less dependent on the sensitive attributes. Since the sensitive attributes are unobserved in training, the method may help us to avoid biases that we are not aware of.

4.2 Treatment-Driven Debiasing

The aim of treatment-driven debiasing is to generate similar representations for inputs with similar non-sensitive attributes but different sensitive attributes. Unlike impact-driven debiasing that mainly decorrelate sensitive information from models for group-level fairness, treatment-driven debiasing focuses on fairness modeling at the individual-level. The existing treatment-driven debiasing methods mainly target fairness issues from individual comparison. for example, “Is it fair that A is chosen but not B?". This is often achieved in a “forward" way with a contrastive learning framework, where the model is trained to generate similar and dissimilar representations for the positive and negative sample pairs, respectively, to reduce the effect of sensitive attributes. The positive sample pairs are often the ones with the same sensitive attributes but different non-sensitive attributes, while the negative sample pairs often have the same or similar non-sensitive attributes but different sensitive attributes. The positive and negative samples can be either selected from the training data or constructed with data augmentation.

A representative work leverages data augmentation and contrastive learning to debias pre-trained text encoders [19]. Specifically, for each sentence in the training data, another sentence with the same semantic meaning but in a different bias direction is generated. Then a contrastive learning loss is used to maximize the mutual information between the representations of two sentences. Additionally, a regularizer is used to minimize the mutual information between the obtained embedding and the sensitive words. Another example uses conditional contrastive learning [79] to remove the effect of sensitive attributes in self-supervised representations. Specifically, the goal of conditional contrastive learning is to obtain similar and dissimilar representations for conditionally-correlated and conditionally-unrelated data, respectively. Here, the condition refers to sensitive attributes. We consider an example of removing gender information from the learned representations. Assuming that we condition on female, the positive pair are data from a female and the corresponding representation from the same female, and the negative pairs are the data from a female and the representations of another female. The positive and negative samples for males can be chosen in a similar fashion. The representations trained in this way will only learn to distinguish the inputs from the same gender but not across genders, which could exclude the gender information from the learned representations. The debiased representations can then be used in other downstream tasks without the undesirable sensitive attributes information.

Some studies have designed contrastive learning strategies tailored for other data types and tasks, such as graph data [53] and candidate generation tasks in recommender systems [89]. In graph data, nodes with similar attributes (including sensitive attributes) tend to be connected. As a result, the graph representation learning methods that exploit the topological structures tend to amplify the bias. To reduce this undesirable effect, [53]

proposes a fairness-aware data augmentation framework, including a feature masking strategy which assigns a larger masking probability to the features that are correlated with the sensitive attributes, and an edge deletion strategy based on whether the sensitive attributes are the same between neighboring nodes. The proposed contrastive learning loss could mitigate the propagation of bias in the graph. Though not targeting sensitive attributes, [89] proposes a contrastive learning strategy to mitigate bias against the under-recommended products for candidates generation in large-scale recommender systems, where the goal is to retrieve a small set of entities from a large corpus. They implement a fixed-size first-in-first-out queue to accumulate positive and negative samples for contrastive learning. It is shown that the model trained with the contrastive loss achieves better fairness on the under-recommended products.

4.3 Hybrid Methods for Implicit Bias Mitigation

Some studies have investigated hybrid methods that address both disparate impact and disparate treatment. A representative line of work is termed *compositional fairness*, where the model is expected to flexibly accommodate different subsets of sensitive attributes, addressing the fairness issues at the subgroup-level.

One strategy to achieve compositional fairness is disentangled representation learning. As we discussed in Section 4.1.3, the key idea of disentangled representation learning is to extract sensitive information out of the representation so that predictors trained on the non-sensitive latent factors are less biased [37]. One advantage of disentangled representation learning is that it can cope with multiple sensitive attributes, and a subset of sensitive attributes can be selected for inference to achieve compositional fairness. Follow-up research has designed more fine-grained disentangled strategies by additionally modeling the intersected information between sensitive and non-sensitive attributes [65] and applied the technique to specific applications, such as facial attribute classification for the hospital no-show [16].

Adversarial learning is an alternative way to achieve compositional fairness. This strategy has been explored on graph data to learn fair graph embeddings [15]. The key idea is to train multiple filters to refine the graph embeddings without particular sensitive attributes, where each filter corresponds to one sensitive attribute. For each filter, a discriminator is defined to predict the corresponding sensitive attribute from the filtered node embeddings. The discriminator is trained with an adversarial loss so that the filter is learned to produce unbiased representations. This design allows us to flexibly apply combinations of the filters to achieve compositional fairness. Although this paper focuses on graph embedding, the idea can be easily generalized and applied to other data types as well.

5 DISCUSSIONS

Prior work has approached machine learning fairness in different directions and proposed various in-processing unfairness mitigation techniques. It is important for the researchers and practitioners to understand the differences among the existing methods. In this section, we provide general discussions on how to choose an appropriate algorithm in real-world applications. We mainly focus on the application scenarios that different methods can tackle and discuss from the perspectives of fairness measurements, task scenarios, base machine learning models, and predictive performance.

The selection of fairness measurements is often the first decision we need to consider since it defines the fairness goals. This is particularly important for the explicit methods, where constraints and regularization terms are designed according to the used measurements. However, due to the subjectivity of fairness, there is no universal agreement about which measurement is better. In practice, some fairness measurements can be incompatible. For example, demographic parity demands each sensitive group to have an equal percentage of positive outcomes, while overall accuracy equality requires an equal accuracy. These two measurements will be fundamentally opposed if different sensitive groups have very different positive rates in the ground truth.

In this case, if both of these two measurements are used, then the unfairness can never be fully eliminated. From a broader context, the appropriate fairness measurements depend on application scenarios, e.g., what legal restrictions are in place, which measures apply, and who gets to decide the relevant measures (e.g., researchers, business owners, or all people affected by the predictions and decisions), which should be discussed from the views of other disciplines, such as laws and social science. From the computational perspective and for machine learning research purposes, one who is only interested in the algorithm design may assume that the suitable measurements are given.

The task scenarios can often imply which mitigation algorithms are suitable. First, different tasks often have different data characteristics and training schemes. A notable example of the impact of data is graph [22], which describes complex relationships and interactions among nodes. A recent study shows that graph structure can amplify unfairness through message passing [44]. Thus, to achieve fairness in graph data, we need a tailored design to counter such amplification, which is often not considered in the mitigation algorithms built on tabular data. Second, different training schemes will often impact the algorithm design. For instance, semi-supervised and unsupervised settings demand tailored designs to leverage unlabeled data to mitigate unfairness [75, 88]. Thus, the algorithm design highly depends on the tasks at hand. Third, due to privacy reasons, we may not have access to sensitive attributes in some applications. If the sensitive attributes are unfortunately not available in a task, then most of the existing mitigation methods will be inapplicable. We instead need to resort to the algorithms that do not require sensitive attributes [38, 40, 55].

The base machine learning models, which often vary in different application scenarios, can put constraints on what mitigation algorithms can be adopted. Firstly, implicit methods can only be applied to deep learning models but not traditional machine learning models since it focuses on debiasing the representations. Thus, for traditional machine models, we often need to design explicit constraints or regularization terms, or resort to pre-processing or post-processing solutions. Secondly, many already complex machine learning systems, such as the recommender system deployed at Facebook [1], often do not allow one to make significant changes to the model. In this case, some mitigation techniques, such as disentangled representation learning, may require broad changes in the model design and be hard to adopt. We refer interested readers to [28, 71] for detailed discussions of how to achieve fairness with minimum modifications on the batch sampling [71] and decoders [28]. However, we note that in some cases (e.g., using any demographic information would be considered unfair), such minimal modifications may not be acceptable. Instead, fairness needs to be considered throughout the entirety of a machine learning process, and applying these techniques needs to be a deliberate choice based on the real situations.

Finally, fairness will influence the predictive performance of machine learning models. However, in this survey, we will not discuss the quantitative relation between fairness and predictive performance since it is often subject to datasets, algorithms, application domains, and whether the hyperparameters are well-tuned. We believe we can not have an affirmative answer without a comprehensive and fair benchmarking of the existing methods. It is worth noting that the relation between predictive performance and fairness is controversial in the existing work. While most studies found that their relationship is a trade-off [21, 33, 46], some argued that they are sometimes in accord [83]. Readers who are interested in fairness and predictive performances and their relationship may refer to [83].

6 RESEARCH CHALLENGES

In-processing methods for machine learning fairness have attracted increasing attention in the community and received significant progresses in recent years. However, there are still several remaining research challenges to be investigated.

Fairness with Missing Sensitive Attributes. While most of the existing studies assume that the sensitive attributes information is available in the dataset, sensitive attributes may not be disclosed or available in many real-world applications. Some laws have imposed restrictions on the use of sensitive information, e.g., the General Data Protection Regulation (GDPR) explicitly requires businesses to protect the personal data for citizens of European Union. Thus, it is a crucial problem to address fairness issues with missing sensitive attributes information. How to achieve fairness in machine learning models remains to be a challenge if very few or even no individual's sensitive attribute is known. Research in this direction is relatively lacking in the literature [40, 55, 74]. More efforts are needed to investigate how to achieve fairness with missing or limited sensitive attributes information.

Fairness with Multiple Sensitive Attributes. The current in-processing techniques mainly focus on addressing the unfairness induced by one sensitive attribute. When multiple sensitive attributes exist, the model that is fair for one sensitive attribute could still be unfair for other sensitive attributes. For instance, both gender and race could lead to unfairness in the *Adult* dataset. The fair model that is designed solely for women will still be likely to be unfair for black women. Additionally, in many real-world scenarios, it is necessary yet challenging to design a model that can flexibly accommodate different fairness needs. For example, we could only require the model to be fair towards gender while we may need to simultaneously consider gender, race, and age in other scenarios. However, the model trained solely for mitigating gender unfairness may not perform well when simultaneously considering gender, race, and age, and vice versa. This practical problem, i.e., how we can accommodate different combinations of sensitive attributes to achieve fairness in different subgroups, is relatively less studied [15, 48, 49] and needs more future work to explore.

Fairness Measurements Selection. Since the design of the mitigation techniques is often inspired by the desired fairness measurements, selecting an appropriate fairness measurement is crucial and depends on the situations at hand. For example, the input of adversarial debiasing [87] depends on the selected fairness measurement because different information is needed for different measurement metrics. Specifically, the prediction $\hat{y}$ is used to calculate demographic parity, whereas both the prediction $\hat{y}$ and the ground truth y are need for equalized odds. It is very likely that a fair model in terms of one measurement remains unfair when evaluated with a different measurement. It is a practical need to understand the effect of fairness measurement on algorithm design to achieve fairness with respect to a single measurement or a combination of multiple measurements. This can help the practitioners to achieve the desired fairness outcome in real-world deployments.

Balance between Group and Individual Fairness. Due to the different goals of group and individual fairness, the existing mitigation techniques often only focus on one of them. However, the model that is optimized to achieve group-level fairness may not be fair in terms of individual-level fairness, and vice versus. In some certain application scenarios, it is desirable to achieve both group and individual fairness. For example, when ranking the items for users in recommender systems, we may aim to minimize the discrepancy between the rankings of each pair of individuals with different sensitive attributes while preserving the overall fairness in the group-level. Such goal could be formulated as a constraint optimization problem, e.g., optimizing individual-level fairness under group-level fairness constraints [35]. More studies are encouraged to investigate the relationships between the group- and individual-level fairness and the techniques to achieve both of them.

Integration of Interpretation and Fairness. Despite the success of machine learning models, they are often been criticized to be uninterpretable. Various interpretable machine learning techniques have been proposed to better understand and debug the model [27], which could serve as the tools to detect and mitigate bias to achieve fairness [29]. For example, in sentiment analysis, local interpretation could detect race biases by getting the feature importance for all the features [51]. Also, local interpretation could serve as regularization terms to achieve fairness by enforcing the predictions to be less dependent on the sensitive attributes [57, 72]. Many interpretation techniques could be potentially adapted to achieve machine learning fairness since the interpretation on features expose the effect of sensitive attributes on the decision-making process. Despite the promise in leveraging this relationship to achieve fairness, interpretation itself remains to be a challenge since it may trigger the artifacts

without careful design [27]. More work on investigating interpretation and particularly the relationship between interpretation and fairness will facilitate the understanding of fairness and motivate the design of unfairness mitigation algorithms.

Relationship between Fairness and Model Performance. Intuitively, if we regard the fairness goals as constraints that limit the decision space of the machine learning model, fairness goals will clearly make the model performance suffer because the resulting set of the possible decision spaces is a subset of the original one. A consensus achieved in most previous work is that the relationship between fairness and model performance is trade-off [21, 33, 46]. However, a recent study reveals that fairness and accuracy are sometimes in accord, e.g., in some semi-supervised tasks [83]. Additionally, due to the subjectivity of fairness, different fairness goals may affect or be affected by model performance differently. As the previous work mainly focuses on empirical studies, a systematic and theoretical investigation of the relationship between fairness and model performance, particularly, under what conditions (such as which fairness goals), improving fairness could also improve model performance, is an important future direction.

Datasets and Benchmarks for Machine Learning Fairness. Benchmark datasets that exhibit discrimination against certain groups of people are what propels us forward in designing unfairness mitigation algorithms. However, benchmark datasets are lacking in the community. The existing studies mainly perform experiments and analyses on a very limited number of small-scale datasets. More efforts in constructing datasets, especially large-scale datasets with demographic information, are encouraged to enable a full exposure of fairness problems. Efforts on benchmarking the existing algorithms under different fairness measurements will also facilitate the research.

7 CONCLUSIONS

In this survey, we present an overview of the current progress of in-processing techniques which focus on modeling design for improving machine learning fairness. Specifically, we categorize the existing techniques into explicit and implicit mitigation methods based on where the fairness issues are tackled in the model. Furthermore, we summarize how each category of methods could be used to mitigate disparate impact (group-level bias), disparate treatment (individual bias), and other mixed scenarios. Finally, we introduce the remaining challenges to be addressed for future research efforts, advocating deeper understandings of fairness and the development of datasets and benchmarks to facilitate future methods design.

ACKNOWLEDGEMENTS

This work is in part supported by NSF grants IIS-1939716 and IIS-1900990. The authors would like to thank Dr. Xia Hu and Dr. Mengnan Du for their constructive feedback. The authors appreciate the constructive feedback from the anonymous reviewers to help improve the paper.

REFERENCES

[1] Bilge Acun, Matthew Murphy, Xiaodong Wang, Jade Nie, Carole-Jean Wu, and Kim Hazelwood. 2021. Understanding training efficiency of deep learning recommendation models at scale. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 802–814.

[2] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. 2018. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*. PMLR, 50–59.

[3] 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.

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

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23, 2016 (2016), 139–159.

[6] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *International Conference on Machine Learning*. PMLR, 405–413.
[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.
[8] 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).
[9] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.
[10] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.
[11] 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 *AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.
[12] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.
[13] Sumon Biswas and Hridesh Rajan. 2021. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM.
[14] Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K D'Mello. 2021. Integrating Psychometrics and Computing Perspectives on Bias and Fairness in Affective Computing: A case study of automated video interviews. *IEEE Signal Processing Magazine* 38, 6 (2021), 84–95.
[15] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*. PMLR, 715–724.
[16] Sabri Boughorbel, Fethi Jarray, and Abdou Kadri. 2021. Fairness in TabNet Model by Disentangled Representation for the Prediction of Hospital No-Show. *arXiv preprint arXiv:2103.04048* (2021).
[17] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, Vol. 2019. NIH Public Access, 2060.
[18] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).
[19] Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders. In *International Conference on Learning Representations*.
[20] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. *Advances in Neural Information Processing Systems* 30 (2017).
[21] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[22] Diane J Cook and Lawrence B Holder. 2006. *Mining graph data*. John Wiley & Sons.
[23] 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.
[24] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120* (2017).
[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
[26] Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. 2020. Counterfactual fairness: removing direct effects through regularization. *arXiv preprint arXiv:2002.10774* (2020).
[27] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.
[28] 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).
[29] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* (2020).
[30] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*. 214–226.
[31] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).
[32] Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640* (2018).

[33] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.
[34] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *IEEE International Conference on Data Engineering*. IEEE, 1918–1921.
[35] David Garcia-Soriano and Francesco Bonchi. 2021. Maxmin-Fair Ranking: Individual Fairness under Group-Fairness Constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
[36] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *AAAI/ACM Conference on AI, Ethics, and Society*. 219–226.
[37] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2021. Counterfactual Fairness with Disentangled Causal Effect Variational Autoencoder. In *AAAI Conference on Artificial Intelligence*.
[38] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv preprint arXiv:1806.11212* (2018).
[39] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).
[40] 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.
[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[42] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *International Conference on World Wide Web*. 173–182.
[43] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*. PMLR, 862–872.
[44] Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, and Xia Hu. 2022. FMP: Toward Fair Graph Message Passing against Topology Bias. *arXiv preprint arXiv:2202.04187* (2022).
[45] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. 2020. Fair prediction with endogenous behavior. In *ACM Conference on Economics and Computation*. 677–678.
[46] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.
[47] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *IEEE International Conference on Data Mining Workshops*. IEEE, 643–650.
[48] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.
[49] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Conference on Fairness, Accountability, and Transparency*. 100–109.
[50] Michael P Kim, Omer Reingold, and Guy N Rothblum. 2018. Fairness through computationally-bounded awareness. In *International Conference on Neural Information Processing Systems*.
[51] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).
[52] Ari Kobren, Barna Saha, and Andrew McCallum. 2019. Paper matching with local fairness constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1247–1257.
[53] Öykü Deniz Köse and Yanning Shen. 2021. Fairness-Aware Node Representation Learning. *arXiv preprint arXiv:2106.05391* (2021).
[54] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).
[55] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *International Conference on Neural Information Processing Systems*.
[56] Peizhao Li, Han Zhao, and Hongfu Liu. 2020. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9070–9079.
[57] Frederick Liu and Besim Avci. 2019. Incorporating Priors with Feature Attribution on Text Classification. In *Annual Meeting of the Association for Computational Linguistics*. 6274–6283.
[58] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019. On the Fairness of Disentangled Representations. In *Advances in Neural Information Processing Systems*, Vol. 32. 14611–14624.
[59] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *ACM SIGKDD international conference on Knowledge discovery and data mining*. 502–510.
[60] 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.
[61] 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.

[62] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research* 9, 5 (2002), 583–597.
[63] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, Vol. 32.
[64] Alejandro Noriega-Campero, Michiel A Bakker, Bernardo Garcia-Bulle, and Alex'Sandy' Pentland. 2019. Active fairness in algorithmic decision making. In *AAAI/ACM Conference on AI, Ethics, and Society*. 77–83.
[65] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2021. Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment. In *AAAI Conference on Artificial Intelligence*, Vol. 35. 2403–2411.
[66] Judea Pearl. 2009. *Causality*. Cambridge university press.
[67] Dana Pessach and Erez Shmueli. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784* (2020).
[68] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *International Conference on Neural Information Processing Systems*.
[69] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *conference on fairness, accountability, and transparency*. 469–481.
[70] John Rawls. 2001. *Justice as fairness: A restatement*. Harvard University Press.
[71] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2020. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696* (2020).
[72] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2007. Right for the right reasons: Training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence*.
[73] Cropanzano Russell. 2001. Three roads to organizational justice. (2001).
[74] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).
[75] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. 2018. Fair coresets and streaming algorithms for fair k-means clustering. *arXiv preprint arXiv:1812.10854* (2018).
[76] Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. 2016. Disciplined convex-concave programming. In *IEEE Conference on Decision and Control*. IEEE, 1009–1014.
[77] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence* (2020).
[78] Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Conference on Fairness, Accountability, and Transparency*. 359–368.
[79] Yao-Hung Hubert Tsai, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Conditional Contrastive Learning: Removing Undesirable Information in Self-Supervised Representations. *arXiv preprint arXiv:2106.02866* (2021).
[80] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).
[81] Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. *arXiv preprint arXiv:2102.12594* (2021).
[82] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE/CVF International Conference on Computer Vision*. 5310–5319.
[83] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: a trade-off revisited. In *International Conference on Neural Information Processing Systems*. 8783–8792.
[84] 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 *International Conference on World Wide Web*. 1171–1180.
[85] 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.
[86] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.
[87] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[88] Tao Zhang, Jing Li, Mengde Han, Wanlei Zhou, Philip Yu, et al. 2020. Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering* (2020).
[89] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3985–3995.