Journal of the American Statistical Association

September 2025 • Volume 120 • Number 551

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Fairness in Machine Learning: A Review for Statisticians

Xianwen He & Yao Li

Fairness in Machine Learning: A Review for Statisticians

Xianwen He and Yao Li*

Department of Statistics & Operations Research, University of North Carolina at Chapel Hill

*yaoli@email.unc.edu

Abstract

With the widespread application of machine learning algorithms in daily life, it is crucial to mitigate the risk of these algorithms producing socially undesirable outcomes that may disproportionately disadvantage certain groups or individuals based on demographic characteristics such as gender, race, or disabilities. In recent years, machine learning fairness has gained increasing attention from both researchers and the public. This article provides a comprehensive overview of fairness-enhancing mechanisms designed to mitigate such risks, along with the fairness criteria they aim to achieve. We organize these fairness-enhancing mechanisms into three categories—pre-processing, in-processing, and post-processing—corresponding to different stages of the machine learning lifecycle and varying levels of access to the underlying algorithm. The discussion focuses on fairness in binary classification models using numerical tabular data, which serve as a foundation for addressing fairness in more complex algorithms. Additionally, we present experimental results that offer a comparative evaluation of representative fairness-enhancing approaches.

Keywords: Classification, Fairness learning, Machine learning ethics, Statistical computing

1 Introduction

1.1 Background

The past decade has witnessed remarkable advancements in machine learning algorithms, which are now incorporated in various aspects of our daily life, such as education (Kučak et al., 2018; Luan and Tsai, 2021), labor market (Faliagka et al., 2012; Pessach et al., 2020), medicine (Chen et al., 2021; Shehab et al., 2022), entertainment (Liao et al., 2024), and legal systems (Northpointe Inc., 2010; Kovalchuk et al., 2024). However, despite their ubiquity, concerns arise regarding potential discrimination embedded in the decisions made by these algorithms (Zemel et al., 2013; Chouldechova, 2017; Karkkainen and Joo, 2021; Mitchell et al., 2021; Fabris et al., 2022; Gao et al., 2023; Pfeiffer et al., 2023). A prominent and troubling example is the recidivism prediction algorithm used in the U.S. criminal justice system (FiveThirtyEight, 2015), which has been criticized for inaccurately predicting higher recidivism rates for African Americans compared to white individuals (Angwin et al., 2016). Other examples include lower prediction accuracy for darker-skinned females in commercial Application Programming Interfaces (APIs) (Buolamwini and Gebru, 2018; Karkkainen and Joo, 2021) for facial recognition from Microsoft (Microsoft Cognitive Services, 2017), IBM (IBM Watson Services, 2017), and Face++ (Megvii Inc., 2017); under-representation and less professional portrayals of women in web search results (Kay et al., 2015); and disparities in ChatGPT responses based on user identity (Eloundou et al., 2024). These cases highlight the risks of deploying machine learning algorithms without fairness guarantees.

Extensive research has been conducted to investigate the causes (Chouldechova and Roth, 2018; Mehrabi et al., 2021; Mitchell et al., 2021), definitions (Zemel et al., 2013; Zafar et al., 2017a; Kusner et al., 2017; Verma and Rubin, 2018), and mitigation techniques (Kamiran et al., 2012; Lohia et al., 2019; Petersen et al., 2021; Wan et al., 2023; Berk et al., 2024) associated with discrimination and biases (often collectively referred to as unfairness) in machine learning algorithms. While fairness is inherently a philosophical concept, a fair machine learning algorithm is generally understood as one that makes decisions independent of an individual's or population's demographic background, and focuses on task-related qualities instead (Chouldechova and Roth, 2018; Mehrabi et al., 2021). Specifically, consider a variable that encodes demographic characteristics such as gender, race, or disability status, which is commonly referred to as a sensitive variable. At a minimum, a fair algorithm must ensure that its decisions do not rely on the sensitive variable, whether directly or indirectly (Hajian and Domingo-Ferrer, 2012; Kamishima et al., 2012).

With the growing integration of machine learning algorithms into daily life, it is crucial to incorporate fairness considerations into their outcomes. This paper aims to present a comprehensive survey of existing algorithmic fairness metrics and fairness-enhancing techniques, with a focus on the machine learning lifecycle. Our objective is to provide a guided overview of available mechanisms and criteria that promote fairness in machine learning, serving as a benchmark for statisticians seeking to improve the fairness of their work and as a starting point for researchers entering the field. Compared to prior surveys on machine learning fairness, this paper makes the following contributions:

- **Detailed explanations of fairness-enhancing mechanisms:** Existing surveys often provide abstract, comment-style overviews of techniques that enforce fairness in machine learning algorithms (Du et al., 2020; Mehrabi et al., 2021; Mitchell et al., 2021; Berk et al., 2021a;

Caton and Haas, 2024). In contrast, this survey offers a structured overview of existing fairness-enhancing mechanisms, accompanied by detailed explanations of the underlying intuitions and formalizations for the most prominent methods. Additionally, we list the resource requirements of each mechanism and explicitly link them to the fairness metrics they aim to satisfy, providing a clearer guide for both researchers seeking appropriate fairness-enhancing methods and beginners looking for an entry point into various techniques.

- **Standardized mathematical notation:** We adopt consistent mathematical notations throughout the paper to facilitate the understanding of key concepts of machine learning fairness and enable meaningful comparisons between different fairness metrics or fairness-enhancing methods.

- **Experimental benchmark:** We implement a selection of representative fairness-enhancing techniques across different stages, i.e., pre-processing, in-processing, and post-processing methods, within a unified framework.

Despite the variety of machine learning algorithms, this survey focuses on fairness metrics and fairness-enhancing techniques for binary classifiers, the foundation for addressing fairness in more complex algorithms. By default, we consider a scenario where an individual's sensitive demographic information is represented by a binary identifier (e.g., 0 or 1), with the value 1 corresponding to the protected group (often the underprivileged group) and 0 to the unprotected group (often the privileged group). For example, 1 is used to represent "female" and 0 indicates "male," or 1 signifies individuals with darker skin tones while 0 corresponds to those with lighter skin tones.

The remainder of this article is organized as follows: Section 2 introduces commonly used quantitative fairness metrics; Section 3 discusses representative fairness-enhancing techniques applied at each stage of the lifecycle of a binary classifier; Section 4 establishes a benchmark and presents experimental results for selected fairness-enhancing methods across different stages; finally, Section 5 concludes the article and offers insights into future research directions in machine learning fairness.

1.2 Formal Notations

Let $f : \mathcal{X} \times \Theta \to \mathbb{R}$ denote the *target model* under fairness assessment, where Θ is the parameter space and $\mathcal{X}$ is the input space. This survey primarily considers *f* as a binary classifier operating on numerical tabular data, i.e., $\mathcal{X} \subseteq \mathbb{R}^d$ and $f : \mathcal{X} \times \Theta \to [0,1]$. When the parameters are not of primary interest, we simplify the notation to $f : \mathcal{X} \to [0,1]$ for the classifier, and $f(\boldsymbol{x})$ for its output given $\boldsymbol{x} \in \mathcal{X}$.

Let $\boldsymbol{X} \in \mathcal{X}$ denote the variables on which the classifier makes decisions, $Y \in \{0,1\}$ represent the class label, and $A \in \{0,1\}$ indicate the group membership. Individuals with $A = 1$ belong to the protected group, while individuals with $A = 0$ belong to the unprotected group. The following texts refer to $\boldsymbol{X}$ as the *features* (explanatory variables), *Y* as the *label* (response variable), and *A* as the *sensitive variable*. While *A* can be multi-dimensional and captures various demographic characteristics (Zafar et al., 2017b; Creager et al., 2019; Jiang et al., 2020), this survey assumes *A* to be a one-dimensional attribute by default.

Suppose that $D=\{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^{N}$ is the given dataset, where each triplet $(\boldsymbol{x}_i, y_i, a_i)$ represents an observation of $(\boldsymbol{X}, Y, A)$. For each instance, a_i serves as the indicator for its group membership, and *D* can be further divided into two subgroups accordingly:

$$D_0 := \{(\boldsymbol{x}_i, y_i, a_i) \in D \mid a_i = 0\}, \quad D_1 := \{(\boldsymbol{x}_i, y_i, a_i) \in D \mid a_i = 1\}.$$

The classifier's decision is denoted by $\hat{Y}$, with $\hat{y}_i$ representing the prediction corresponding to $(\boldsymbol{x}_i, y_i, a_i)$. Notably, $f(\boldsymbol{x})$ does not return a discrete label but rather the score of a positive outcome. In practice, the predicted label is generally determined as

$$\hat{y}_i = \mathbb{I}\{f(\boldsymbol{x}_i) > \tau\}, \quad i = 1, \ldots, N$$

where $\tau \in [0,1)$ is a given threshold, and it is commonly set to be 0.5.

2 Fairness Definitions and Metrics

This section introduces commonly used fairness metrics and their interrelationships. Intuitively, fairness metrics are designed to measure the influence of sensitive variables on model predictions. Existing fairness notions can be categorized into two main types: group fairness notions and individual fairness notions (Zemel et al., 2013; Calmon et al., 2017; Caton and Haas, 2024). *Group fairness metrics* typically require equitable performance across subgroups defined by the sensitive attributes. In contrast, *individual fairness metrics* focus on ensuring similar outcomes for similar inputs. As discussed in Section 1, this section summarizes and interprets fairness metrics for binary classifiers. It is important to note that we concentrate on metrics reflecting the fairness level of the machine learning algorithms, while there exist metrics designed to assess the fairness of datasets (Kamiran and Calders, 2012). Additionally, this section does not aim to provide an exhaustive list of all existing fairness metrics but rather highlights the most frequently used ones. In practice, it is common to propose task-specific metrics to address the unique requirements of particular applications (Burke et al., 2018; Eloundou et al., 2024).

2.1 Group Fairness

Group fairness notions evaluate the outcomes of a classification algorithm by comparing results across two or more groups defined by sensitive attribute(s). They can be further divided into two sub-categories: parity-based metrics and calibration-based metrics.

2.1.1 Parity-based Metrics

In general, parity-based metrics require the (conditional) distribution of $\hat{Y}$ to be similar to a certain degree across protected and unprotected groups.

Demographic Parity

Demographic Parity, also known as *Statistical Parity* or equivalently, *Disparate Impact*, is oriented by the expectation that "the overall proportion of members in a protected group receiving positive (negative) classification is identical to the proportion of the population as a whole" (Zemel et al., 2013; Zafar et al., 2017a, b; Kusner et al., 2017). A classifier satisfies Demographic Parity, or is free of Disparate Impact, if

$$P(\hat{Y} = y | A = 0) = P(\hat{Y} = y | A = 1), y \in \{0,1\}. \text{ (1)}$$

In practice, the difference between the proportions of receiving positive predictions across the protected and unprotected groups is often applied to measure the magnitude of discrimination of the outcomes (Kamishima et al., 2012; Kamiran and Calders, 2012; Zemel et al., 2013). Therefore, an equivalent expression for Demographic Parity is (Pessach and Shmueli, 2022)

$$| P(\hat{Y} = 1 | A = 1) - P(\hat{Y} = 1 | A = 0) | < \epsilon \text{ (2)}$$

where ϵ is a given positive threshold.

Demographic Parity is a widely accepted quantification of the 80%-rule (or more generally, p%-rule) advocated by the US Equal Employment Opportunity Commission (Biddle, 2017), which states that the ratio between the percentage of members from the protected group receiving positive decisions and the percentage of those from the unprotected group should be no less than *p* over 100:

$$\frac{P(\hat{Y} = 1 | A = 1)}{P(\hat{Y} = 1 | A = 0)} \geq p / 100 \text{ (3)}$$

where $p \in (0,100]$.

By replacing the prediction $\hat{Y}$ with the ground-truth label *Y*, the formula above can be applied to measure the magnitude of fairness/discrimination of the data set *D* (Feldman et al., 2015; Kamiran and Calders, 2012). This survey focuses on algorithmic fairness and considers classification outcomes by default. However, we recommend our readers pay attention to the difference between $\hat{Y}$ and *Y* when determining the suitable fairness notions for their tasks.

Equalized Odds and Equal Opportunity

Proposed by Hardt et al. (2016), *Equalized Odds* and *Equal Opportunity* aim to reach a balance between fairness and prediction accuracy. Compared to Demographic Parity, these two notions emphasize the model utility that requires $\hat{Y}$ to well approximate *Y*. A classifier satisfies Equalized Odds if

$$P(\hat{Y} = 1 | A = 0, Y = y) = P(\hat{Y} = 1 | A = 1, Y = y), \forall y \in \{0,1\}. \text{ (4)}$$

Equal Opportunity is formulated as a relaxation for Equalized Odds. In the binary case, the label $Y = 1$ is often considered as qualifying *advantaged* decisions, such as "not defaulting on a load," "admission to a college," or "receiving a promotion" (Hardt et al., 2016). Equal Opportunity only requires non-discrimination within the group with advantaged labels, which gives rise to its name. A classifier satisfies Equal Opportunity if

$$P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1). \text{ (5)}$$

This survey considers Equalized Odds equivalent to the notion of *Conditional Procedure Accuracy Equality* (Berk et al., 2021b), which ensures that the True Positive Rate (TPR) and True Negative Rate (TNR) are identical for both the protected and unprotected groups. Notably, there exists a subtle distinction between $P(\hat{Y} = y' \mid Y = y)$ and its empirical estimate:

$$\sum_i \mathbb{I}\{\hat{y}_i = y', y_i = y\} / \sum_i \mathbb{I}\{y_i = y\},$$

as well as between $P(\hat{Y} = y' \mid Y = y, A = a)$ and its corresponding empirical form:

$$\sum_i \mathbb{I}\{\hat{y}_i = y', y_i = y, a_i = a\} / \sum_i \mathbb{I}\{y_i = y, a_i = a\}.$$

However, in practice, fairness notions are typically assessed based on the available dataset. Therefore, this survey does not distinguish between conditional probabilities and confusion-matrix-based statistics when defining fairness metrics. To maintain consistency, conditional probabilities are used by default.

Fairness notions that follow a similar intuition to Equalized Odds include *Equalizing Disincentives* (Jung et al., 2020) and *Treatment Equality* (Berk et al., 2021b). Equalizing Disincentives requires that the difference between the TPR and FPR be the same across the protected and unprotected groups:

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) - P(\hat{Y} = 1 \mid Y = 0, A = 0) \quad (6)$$

$$= P(\hat{Y} = 1 \mid Y = 1, A = 1) - P(\hat{Y} = 1 \mid Y = 0, A = 1). \quad (7)$$

Treatment Equality, on the other hand, requires that the ratio of False Negatives (FN) to False Positives (FP) is the same across the protected and unprotected groups:

$$\frac{\sum_i \mathbb{I}(\hat{y}_i = 0, y_i = 1, a_i = 0)}{\sum_i \mathbb{I}(\hat{y}_i = 1, y_i = 0, a_i = 0)} = \frac{\sum_i \mathbb{I}(\hat{y}_i = 0, y_i = 1, a_i = 1)}{\sum_i \mathbb{I}(\hat{y}_i = 1, y_i = 0, a_i = 1)}. \quad (8)$$

It is straightforward to show that a classifier satisfying Equalized Odds will satisfy Equalizing Disincentives; and if the proportion of ground-truth positive labels is identical across the protected and unprotected groups, it will satisfy Treatment Equality as well.

Overall Accuracy Equality

A classifier satisfies *Overall Accuracy Equality* (Berk et al., 2021b) if it achieves the same level of accuracy for the protected and unprotected groups:

$$P(\hat{Y} = Y \mid A = 0) = P(\hat{Y} = Y \mid A = 1). \quad (9)$$

Since accuracy is a fundamental metric reflecting the utility of a machine learning model, Overall Accuracy Equality serves as a straightforward measure of the expectation that the model performs equivalently across subgroups defined by the sensitive attribute.

Disparate Treatment

A classifier suffers *Disparate Treatment* if "its decisions are (partly) based on the subject's sensitive attribute information" (Zafar et al., 2017b). The difference between Disparate Treatment and Disparate Impact lies in the fact that the former takes both the sensitive variable *A* and the innocuous variables $\boldsymbol{X}$ into consideration. A classifier does NOT suffer from Disparate Treatment if (Zafar et al., 2017a)

$$P(\hat{Y}=y|\ \boldsymbol{X},A)=P(\hat{Y}=y|\ \boldsymbol{X}), y\in\{0,1\} \ (10)$$

where $P(\hat{Y}=y|\cdot)=\mathbb{E}(\mathbb{I}\{\hat{Y}=y\}|\cdot)$. A simple way to avoid Disparate Treatment is not to include the sensitive attribute when building the machine learning model.

Conditional Equality of Opportunity

The notion *Conditional Equality of Opportunity* (Beutel et al., 2019) takes one step further from Equalized Odds and Disparate Treatment, considering the probability of receiving positive outcomes conditional on both the sensitive attribute and a chosen feature. In the binary case, Conditional Equality of Opportunity, conditioned on a feature *S*, is defined as

$$P(\hat{Y}=1|\ A=0,Y=0,S=s)=P(\hat{Y}=1|\ A=1,Y=0,S=s), \forall s\in\mathcal{S} \ (11)$$

where *S* is a selected feature valued in $\mathcal{S}$. Furthermore, by weighting each gap by p_s for $s\in\mathcal{S}$, one can control the prioritization of different values of *S* and define the notion of *Conditional Equality of Opportunity Gap* as

$$\text{EOGap}=\sum_{s\in\mathcal{S}} p_s\left[P(\hat{Y}=1|\ A=0,Y=0,S=s)-P(\hat{Y}=1|\ A=1,Y=0,S=s)\right] \ (12)$$

where $p_s\in[0,1]$ and $\sum_{s\in\mathcal{S}} p_s=1$. Setting $p_s=1/|\mathcal{S}|$ assigns equal weight to each value in $\mathcal{S}$, which is generally a reasonable choice in the absence of a strong reason for alternative weighting. However, if a group exhibits a skew in feature *S*, another option is to set $p_s=P(S=s|\ A=1)$, aligning the gap in False Positive Rates (FPRs) with the background distribution of the focused group.

2.1.2 Calibration-based Metrics

Calibration is a crucial condition for risk assessment tools (Kleinberg et al., 2017; Pleiss et al., 2017; Berk et al., 2023). According to Kleinberg et al. (2017), a risk assessment tool consists of a series of "bins" and a mapping that assigns each input instance across multiple bins with certain probabilities. Each bin *b* is labeled with a *score* $v_b\in[0,1]$ representing the decision given to instances within *b*. Such a tool satisfies *calibration within groups* if for each protected/unprotected group and each bin *b*, exactly v_b fraction of the expected number of instances from this group assigned to *b* belongs to the ground-truth positive class. Following this idea, under the binary classification context, a classifier $f:\mathcal{X}\to[0,1]$ is said to be *perfectly calibrated* (Pleiss et al., 2017) if

$$P(Y=1|\ f(\boldsymbol{X})=p, A=a)=p,\ \forall a\in\{0,1\}, p\in[0,1]. \ (13)$$

In general, the essential property of calibration can be interpreted as “ the scores mean what they claim to mean” (Kleinberg et al., 2017), regardless of membership in the protected/unprotected groups. Therefore, the scale of the score can be relaxed, as long as its independence of the sensitive attribute is retained. Suppose that $f:\mathcal{X}\rightarrow\mathbb{R}$ is the score function and $\delta\in\mathbb{R}$ is a given threshold. Instances with $f(\boldsymbol{x})>\delta$ are deemed as high-risk, and the remaining are regarded as low-risk. There are two relaxed calibration-based metrics: *well calibrated* and *Predictive Parity* (Dieterich et al., 2016; Chouldechova, 2017). A score is said to be well calibrated if

$$P(Y=1|\ f(\boldsymbol{X})=s,A=0)=P(Y=1|\ f(\boldsymbol{X})=s,A=1),\ \forall s\in\mathbb{R}\ (14)$$

and a score satisfies Predictive Parity at the threshold δ if

$$P(Y=1|\ f(\boldsymbol{X})>\delta,A=0)=P(Y=1|\ f(\boldsymbol{X})>\delta,A=1).\ (15)$$

In binary classification, if $\hat{Y}=\mathbb{I}\{f(\boldsymbol{X})>\delta\}$, Equation (15) is equivalent to

$$P(Y=1|\ \hat{Y}=1,A=0)=P(Y=1|\ \hat{Y}=1,A=1).\ (16)$$

It is worth noticing that, despite the similarity between Equations (14) and (15), well-calibrated scores can violate Predictive Parity because the conditional distribution $f(\boldsymbol{X})\,|\,A$ may differ across the protected/unprotected groups.

Conditional Use Accuracy Equality (Berk et al., 2021a) and *False Discovery Rate (FDR)* (Wang et al., 2019) are metrics conceptually related to calibration. Conditional Use Accuracy Equality follows a format similar to Equalized Odds but conditions on the predicted outcomes rather than the ground-truth labels:

$$\begin{aligned}&P(Y=1|\ A=0,\hat{Y}=1)=P(Y=1|\ A=1,\hat{Y}=1)\\&P(Y=0|\ A=0,\hat{Y}=0)=P(Y=0|\ A=1,\hat{Y}=0).\end{aligned}\ (17)$$

FDR captures disparities in the rates of incorrect positive predictions between protected and unprotected groups, and is defined as:

$$\mathrm{FDR}=P(Y=0|\ \hat{Y}=1,A=1)-P(Y=0|\ \hat{Y}=1,A=0).\ (18)$$

As previously mentioned, calibration is a critical property for risk assessment tools. A prominent example is the recidivism prediction instrument (RPI), which estimates the likelihood of a criminal defendant reoffending. In such a system, where $Y=1$ indicates recidivism and $Y=0$ denotes non-recidivism, calibration-based metrics help ensure that predicted risk scores are well aligned with actual recidivism probabilities, thereby mitigating unethical outcomes across different groups.

Despite their importance, calibration-based metrics can be incompatible with parity-based metrics (Kleinberg et al., 2017; Chouldechova, 2017). In fact, Kleinberg et al. (2017) demonstrated that, except in highly constrained cases, the following three fairness notions cannot be simultaneously satisfied: calibration within groups, parity of average scores for instances with ground-truth negative labels across groups, and parity of average scores for instances with ground-truth positive labels across groups.

2.2 Individual Fairness

In general, individual fairness notions aim to ensure that similar items should be treated similarly by the model (Zemel et al., 2013). Some references (Dwork et al., 2012; Kusner et al., 2017) refer to such a property as “fairness through awareness.” This survey avoids that terminology, since the seeming counterpart, “fairness through unawareness,” does not refer to the failure in individual fairness, but the property that the sensitive attribute is not explicitly used in the decision-making process (Kusner et al., 2017; Grgić-Hlača et al., 2016).

Compared to group fairness, individual fairness focuses on each participating individual instead of the overall property of certain groups. Therefore, some individual metrics do not explicitly include sensitive variables in their formula, but rely on disparities between outputs for each instance and for its neighbors. Additionally, unlike group fairness, individual fairness notions are often highly task-dependent, because of the design of the metrics for the *similarity* between the input and output. In the following tasks, we will introduce a few representative notions to help the readers understand the concept of individual fairness.

Consistency

Consistency (Zemel et al., 2013) of a given classifier can be formulated as

$$1-\frac{1}{Nk}\sum_{i=1}^{N}|\hat{y}_i - \sum_{j\in \mathrm{kNN}(\boldsymbol{x}_i)} \hat{y}_j \,| \quad (19)$$

where $k \in \{1,2,3,\ldots\}$ and $\mathrm{kNN}(\boldsymbol{x})$ refers to the index set of the *k*-nearest neighbors of $\boldsymbol{x}$, which is obtained by the kNN clustering applied to the full data set. As the name suggests, consistency requires model classifications to be locally stable in the input space, with values closer to 1 indicating a higher degree of individual fairness. It enforces similar outputs for instances with similar features, regardless of their group membership.

Counterfactual Fairness

Counterfactual fairness (Kusner et al., 2017) is designed based on the concept and toolkit of causal inference. Suppose that (U,V,F) is a causal model, where *U* are latent (background) variables, $V = A \cup \boldsymbol{X}$ are observable variables including the sensitive attribute *A*, and *F* is a set of functions defining structural equations such that *V* is a function of *U*. For a binary classifier, counterfactual fairness holds if, for any context where $\boldsymbol{X}=\boldsymbol{x}$ and $A=a$, there is

$$P(\hat{Y}_{A\leftarrow a}(U)=y|\ \boldsymbol{X}=\boldsymbol{x}, A=a) = P(\hat{Y}_{A\leftarrow a'}(U)=y|\ \boldsymbol{X}=\boldsymbol{x}, A=a),\ \forall y, a' \in \{0,1\}.\quad (20)$$

In this equation, the notation $\hat{Y}_{A\leftarrow a}(u)$ stands for “the value of $\hat{Y}$ if *A* had taken value *a* for a given $U=u$,” following the standard formulation of counterfactual quantities in causal inference (Neuberg, 2003). In general, the counterfactual fairness definition is based on the “intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belongs to a different demographic group”

(Kusner et al., 2017). In the Appendices, we will provide a more detailed introduction to this metric along with the relationship between causal inference and machine learning fairness.

Generalized Entropy Index

Generalized Entropy Index (GEI) (Speicher et al., 2018) quantifies algorithmic discrimination by measuring deviations in individual predictions (b_i) from the average prediction accuracy (μ). For a given constant $\alpha \notin \{0,1\}$, GEI is defined as

(Tex translation failed) (21)

where $b_i = \hat{y}_i - y_i + 1$ and $\mu = \sum_i b_i / N$. This notion can be interpreted as a measure of theoretical redundancy for the information contained in data.

2.3 Challenges beyond Algorithmic Fairness Metrics

Bridging the philosophical concept of fairness to a quantitative notion based on the input and output of machine learning algorithms is inherently challenging. Intuitively, algorithmic fairness aims to eliminate subjective discrimination in human decision-making processes. However, the interpretation of discrimination and fairness can be task-dependent. A series of studies have explored the choices, assumptions, and trade-offs among various algorithmic fairness notions (Verma and Rubin, 2018; Mitchell et al., 2021; Pfeiffer et al., 2023; Berk et al., 2023).

A prominent example illustrating the impact of fairness metric choices is the debate over COMPAS (Northpointe Inc., 2010), a risk assessment tool that estimates the likelihood of future criminal behavior. In a widely cited report by ProPublica (Angwin et al., 2016), COMPAS was accused of exhibiting significant racial disparities. Although the tool made mistakes at roughly similar rates across racial groups, it was found to be twice as likely to falsely flag Black defendants as future criminals compared to white defendants. However, this investigation was soon criticized for its choice of fairness metrics. Critics argued that, rather than focusing on False Positive Rates and False Negative Rates, calibration was a more appropriate metric for risk assessment tools, which COMPAS well satisfied (Flores et al., 2016; Dieterich et al., 2016). Later, Kleinberg et al. (2017) formalized and clarified this controversy, proving the general incompatibility between calibration-based metrics and Equalized Odds. This example underscores that an algorithm deemed fair under one criterion can be unfair under another.

The justification of algorithmic fairness begins once the application scenario is identified. It is shaped by the overarching social objective, the population affected by the decision, and the algorithm’s decision space, each of which involves choices that ultimately influence the model’s ethical impact (Mitchell et al., 2021). Even when focusing solely on the algorithm, aspects such as datasets (VanderWeele and Hernán, 2012; Kleinberg et al., 2018) and model selection (Chouldechova and G’Sell, 2017) can significantly affect fairness outcomes. While this survey does not dive deeply into topics concerning ethnicity, sociology, or criminology, it is important to recognize that the choice and definition of fairness metrics depend on the full context of the application scenario. As such, they often require substantial justification and interdisciplinary collaboration.

3 Fairness-Enhancing Mechanisms

This section provides an overview of fairness-enhancing mechanisms designed to mitigate potential discrimination in machine learning algorithms. Following previous literature (Du et al., 2020; Berk et al., 2021a; Pessach and Shmueli, 2022; Caton and Haas, 2024), these methods can be broadly categorized into three groups based on the stage of the machine learning lifecycle in which they are applied: *pre-processing*, *in-processing*, and *post-processing* methods. While no universally rigorous definitions exist for each category, we distinguish these fairness-enhancing mechanisms by their dependence on training data, the target machine learning model, and the training process. Furthermore, we discuss the fairness notions that each method aims to optimize. This perspective is intended to help readers identify the most suitable fairness-enhancing approach based on their access to the model and specific fairness objectives.

3.1 Pre-processing

While pre-processing fairness-enhancing techniques can be interpreted in various ways, this survey defines them as methods that (1) aim to eliminate discrimination in model output and (2) are applied before the implementation of the target classifier. Generally, pre-processing techniques do not require users to modify the training process, access model parameters, or infer well-trained models. Instead, they primarily operate by manipulating the training dataset. Additionally, since these methods function independently of directly modifying or inferring the classifier, further evaluation is often necessary to ensure that the fairness metrics desired during their development are preserved in the model's predictions.

Intuitively, pre-processing techniques seek to remove information related to sensitive attributes from the training dataset. The simplest approach is to exclude the sensitive attribute *A* and its most correlated variables from the dataset, a process known as *blinding* (Kamiran and Calders, 2012). However, this approach has been shown to be insufficient for preventing discrimination and may significantly degrade the model's utility (Kamishima et al., 2012). This phenomenon, referred to as *indirect discrimination* (Pedreshi et al., 2008) or the *red-lining effect* (Calders and Verwer, 2010), arises from the correlations between the sensitive attribute *A* and the non-sensitive attributes $\boldsymbol{X}$, highlighting the need for more nuanced dataset modifications to mitigate discrimination effectively.

More sophisticated and effective pre-processing fairness-enhancing mechanisms can be broadly classified into three categories: modifying the training data distribution, learning fair representations, and decoupling. The following subsections introduce the fundamental ideas and representative methods within each category.

3.1.1 Modifying Training Data Distribution

One common pre-processing technique for eliminating discrimination is to perturb the original training set *D* to generate a new set, $\tilde{D} = \{(\boldsymbol{x}_i, \tilde{y}_i, \tilde{a}_i)\}$, which can be regarded as a finite sample drawn from a discrimination-free distribution over $\mathcal{X} \times \{0,1\} \times \{0,1\}$. Specifically, suppose that $(\boldsymbol{X}, Y, A)$ follows the distribution $\mathcal{D}$, which is potentially biased, and the training set *D* is a finite sample of size *N* drawn from $\mathcal{D}$. Techniques under this category assume that there exists a non-discriminatory distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\} \times \{0,1\}$ and aim to find a sample $\tilde{D}$ drawn from $\mathcal{D}$ based on the observed values D. Specific approaches to perturb the training set include sampling, reweighting, relabeling (data massaging), and feature perturbation.

Reweighting and Sampling

A straightforward approach to perturbing the data distribution is to assign customized weights to instances in the dataset (Calders et al., 2009; Kamiran and Calders, 2012; Krasanakis et al., 2018; Jiang and Nachum, 2020). The key step in reweighting is to carefully select weights that compensate for biases embedded in the original input. In practice, these weights are typically derived from the discrepancy between observed values and an ideal, discrimination-free distribution estimated based on the desired notions of fairness.

Kamiran and Calders (2012) proposed a naive approach for determining instance weights. Suppose P_{fair} represents the ideal joint distribution ensuring independence between Y and A, then

$$P_{\text{fair}}(Y=y, A=a) = P(A=a)P(Y=y), \quad \forall a, y \in \{0,1\}.$$

The estimated ideal joint distribution, $\hat{P}_{\text{fair}}$, can be computed as

$$\hat{P}_{\text{fair}}(Y=y, A=a) = \frac{\sum_{i=1}^{N} \mathbb{I}\{a_i = a\}}{N} \times \frac{\sum_{i=1}^{N} \mathbb{I}\{y_i = y\}}{N}.$$

Conversely, based on the observed dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, the empirical joint distribution of Y and A is estimated as

$$\hat{P}(Y=y, A=a) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{y_i = y, a_i = a\}.$$

To compensate for the difference between $\hat{P}_{\text{fair}}$ and $\hat{P}$, the weight w_i for each instance $(\boldsymbol{x}_i, y_i)$ with the sensitive indicator a_i is naturally derived as

$$w_i = \frac{\hat{P}_{\text{fair}}(Y=y_i, A=a_i)}{\hat{P}(Y=y_i, A=a_i)}. \quad (22)$$

Sampling is a similar technique to reweighting (Kamiran and Calders, 2012; Shekhar et al., 2021). It aims to build a discrimination-free dataset by sampling accordingly from the protected and unprotected groups of the original dataset. In fact, sampling is equivalent to count-based reweighting methods but can be applied for classifiers that do not work directly with weights (Kamiran and Calders, 2012).

In recent years, more refined reweighting and sampling methods have been proposed. For instance, Jiang and Nachum (2020) introduced a framework that designed instance weights to ensure fairness in label distributions according to specified group fairness metrics. They assumed the existence of an unbiased label Y_{true}, which had been overwritten by a biased agent as Y_{bias}. The intuition is to recover Y_{true} by reweighting the given set *D*, which consists of observations of Y_{bias}.

In this case, when $\boldsymbol{X}$ is given, Y_{true} and Y_{bias} follow Bernoulli distributions determined by $\boldsymbol{X}$ respectively. Suppose that $h:\mathcal{X}\to[0,1]$ is an arbitrary probability label function, and $\boldsymbol{X}$ follows the distribution $\mathcal{P}_{\boldsymbol{X}}$. Together they define a random variable *Y* that, given $\boldsymbol{X}=\boldsymbol{x}$, follows a Bernoulli distribution taking the value 1 with probability $h(\boldsymbol{x})$. Let h_{true} be the probability label function for Y_{true}, and h_{bias} be the label function for Y_{bias}:

$$h_{\text{true}}(\boldsymbol{x}) := P(Y_{\text{true}}=1 \mid \boldsymbol{X}=\boldsymbol{x}), h_{\text{bias}}(\boldsymbol{x}) := P(Y_{\text{bias}}=1 \mid \boldsymbol{X}=\boldsymbol{x}).$$

Then the objective is to recover h_{true} by h_{bias}.

The authors adopted a general notion for group fairness metrics (Goh et al., 2016; Cotter et al., 2019). Given $\boldsymbol{X}\sim\mathcal{P}_{\boldsymbol{X}}$ and $Y \mid \boldsymbol{X} \sim \text{Bernoulli}(h(\boldsymbol{X}))$, many group fairness metrics for the function *h* can be expressed by

$$\mathbb{E}_{\boldsymbol{X}\sim\mathcal{P}_{\boldsymbol{X}}}[\mathbb{E}_{Y\sim\text{Bern}(h(\boldsymbol{X}))}c(\boldsymbol{X},Y)] = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{P}_{\boldsymbol{X}}}[h(\boldsymbol{X})c(\boldsymbol{X},1)+(1-h(\boldsymbol{X}))c(\boldsymbol{X},0)] = 0 \quad (23)$$

where $c:\mathcal{X}\times\{0,1\}\to\mathbb{R}$ characterizes a specific fairness constraint. For example, setting $c(\boldsymbol{x},0)=0$ and

$$c(\boldsymbol{x},1) = \frac{\mathbb{I}(A=1,\boldsymbol{X}=\boldsymbol{x})}{P(A=1)} - 1$$

yields Demographic Parity, i.e., $P(Y=1,A=1)=P(Y=1)P(A=1)$. Similarly, Equal Opportunity and Equalized Odds can be derived by adjusting $c(\cdot,\cdot)$. While a rigorous definition of *c* should incorporate $\boldsymbol{X}$, *Y*, the sensitive attribute *A*, and sometimes Y_{true}, for simplicity, we denote it as $c(\boldsymbol{X},Y)$ in this section. It is worth noticing that the model *f* is also a probability label function, and thereby can be measured by the metric (23) as well.

Following this general notion, if h_{true} is unbiased with respect to the given fairness metric *c*, it is natural to have

$$\mathbb{E}_{\boldsymbol{X}\sim\mathcal{P}_{\boldsymbol{X}}}[\mathbb{E}_{Y\sim\text{Bern}(h_{\text{true}}(\boldsymbol{X}))}(c(\boldsymbol{X},Y))] = 0.$$

Intuitively, h_{bias} can be regarded as a function similar to h_{true} but tolerating a certain level of bias. Therefore, if h_{true} satisfies fairness metrics $c_1,\ldots,c_K$, then for some $\epsilon_1,\ldots,\epsilon_K\in\mathbb{R}$, h_{bias} solves the following constrained optimization problem

$$\begin{aligned}&\arg\min_{h:\mathcal{X}\to[0,1]} \mathbb{E}_{\boldsymbol{X}\sim\mathcal{P}_X}[D_{KL}(h(\boldsymbol{X})\,\|\,h_{\text{true}}(\boldsymbol{X}))]\\&\text{s.t.}\,\mathbb{E}_{\boldsymbol{X}\sim\mathcal{P}_X}[\mathbb{E}_{Y\sim\text{Bern}(h(\boldsymbol{X}))}[c_k(\boldsymbol{X},Y)]]=\epsilon_k,\quad \forall k=1,\ldots,K\end{aligned}$$

where D_{KL} represents the KL-divergence, and the closed-form solution is

$$h_{\text{bias}}(\boldsymbol{x}) \propto h_{\text{true}}(\boldsymbol{x})\cdot\exp\{-\sum_{k=1}^{K}\lambda_k c_k(\boldsymbol{x},1)\}.$$

Accordingly, the unbiased label is expressed in terms of the biased label as

$$h_{\text{true}}(\boldsymbol{x}) \propto h_{\text{bias}}(\boldsymbol{x})\cdot\exp\{\sum_{k=1}^{K}\lambda_k c_k(\boldsymbol{x},1)\},\ \text{for some}\ \lambda_1,\ldots,\lambda_K\in\mathbb{R}.\ (24)$$

To obtain the reweighting parameters, first define

$$\tilde{w}_i = \exp(\sum_{k=1}^{K}\lambda_k\cdot\mathbb{I}[\boldsymbol{X}=\boldsymbol{x}_i, A=1]),\quad \forall i=1,\ldots N\ \ (25)$$

and then the weight for instance *i* is calculated by

$$w_i=\begin{cases}\tilde{w}_i/(\tilde{w}_i+1), \text{if } y_i=1,\\ 1/(\tilde{w}_i+1), \text{if } y_i=0.\end{cases}\ (26)$$

In practice, λ_i can be learned iteratively along with the target classifier *f*. Let f_t be the model trained on the given set *D* with learned bias-correcting weights $\{w_i^t\}_{i=1}^N$ at iteration *t*. Define

$$\Delta_k^t := \mathbb{E}_{\boldsymbol{X}\sim\mathcal{P}_X}[\mathbb{E}_{Y\sim\text{Bern}(f_t(\boldsymbol{X}))}(c_k(\boldsymbol{X},Y))],\quad \forall k=1,\ldots,K$$

and the coefficients $\lambda_1^{t+1},\ldots,\lambda_K^{t+1}$ for the iteration $t+1$ can be obtained by

$$\lambda_k^{t+1}\leftarrow\lambda_k^t-\eta\Delta_k,\quad k=1,\ldots K.$$

Then $\{w_i^{t+1}\}_{i=1}^N$ can be calculated according to Equation (25) and (26), and f_{t+1} is trained with these updated weights. It is worth noticing that such an implementation is not strictly pre-processing because of its interaction with the classifier's training process. However, Formula (24) suggests that the Y_{true} could potentially be recovered independently of the target model. Therefore, this paper still categorizes this reweighting technique as a pre-processing method.

Relabeling and Feature Perturbation

Another approach to perturb the training set distribution is to directly modify the values of *Y* for certain instances. This strategy is known as relabeling or data massaging (Hajian and Domingo-Ferrer, 2012; Kamiran and Calders, 2012). In scenarios where members of the protected group are less likely to receive positive decisions compared to members with similar features from the unprotected group, the fundamental idea of relabeling is to flip some

y_i in the protected group from 0 to 1 while adjusting the same number of y_j in the unprotected group from 1 to 0 (Kamiran and Calders, 2012). The essential step of relabeling is to determine how many labels to flip and which data instances should be altered. Recall that D_0 and D_1 represent the unprotected and protected groups, respectively. We further define

$$D_0^+ := \{(\boldsymbol{x}_i, y_i, a_i) \in D |\ a_i = 0, y_i = 1\}, \quad D_1^- := \{(\boldsymbol{x}_i, y_i, a_i) \in D |\ a_i = 1, y_i = 0\}.$$

Then a naive relabeling approach (Kamiran and Calders, 2012) for enhancing Demographic Parity is to define the number of labels to flip in each group as

$$M := \left| \frac{|D_0^+|}{|D_0|} - \frac{|D_1^-|}{|D_1|} \right| \times \frac{|D_0| \times |D_1|}{|D|}$$

where *M* represents the number of instances from D_1^- that will be flipped from a negative label to a positive label, and the number of instances from D_0^+ that will be flipped from a positive label to a negative label. The intuition behind this approach is to increase the number of positive labels in the protected group while decreasing the number of positive labels in the unprotected group, thereby mitigating the discrepancy between them. A ranker $\mathcal{R}$ is typically applied to select the instances to be flipped. When the score $\mathcal{R}(\boldsymbol{x})$ is consistent with the probability of receiving a positive label, the top-*M* scoring instances from D_1^- and the lowest-*M* scoring instances from D_0^+ are selected for relabeling.

An alternative to relabeling is to modify the value of *A* instead of *Y* (Hajian and Domingo-Ferrer, 2012). For example, Luong et al. (2011) proposed a method to evaluate individual fairness for each member of the protected group by comparing their label with that of their unprotected counterpart. The underlying intuition is that instances with similar, legally admissible characteristics, regardless of whether they belong to the protected group, should receive similar decisions. Specifically, given an instance $(\boldsymbol{x}, y, a)$ with the sensitive indicator $a = 1$, one can create a counterpart instance by flipping *a* from 1 to 0 while keeping $\boldsymbol{x}$ unchanged. The label of this counterpart is determined based on its *k*-nearest neighbors within the unprotected group. The discrepancy between *y* and the counterpart's label reflects the potential discrimination against the instance $(\boldsymbol{x}, y, a)$. To formally address this measure, suppose that there exists a distance metric $\mathrm{dist}(\cdot,\cdot)$ on $\mathcal{X}$. Then let $\mathcal{K}_0(\boldsymbol{x})$ be the set of *k*-nearest neighbors of $\boldsymbol{x}$ within $D_0|_{\mathcal{X}} = \{\boldsymbol{x}_i |\ (\boldsymbol{x}_i, y_i, a_i) \in D_0\}$, and similarly, let $\mathcal{K}_1(\boldsymbol{x})$ be the set of *k*-nearest neighbors within $D_1|_{\mathcal{X}} = \{\boldsymbol{x}_i |\ (\boldsymbol{x}_i, y_i, a_i) \in D_1\}$. The discrepancy measure is defined as

$$\mathrm{diff}(\boldsymbol{x}) = \frac{1}{k}\Big(|\{\boldsymbol{x}' \in \mathcal{K}_1(\boldsymbol{x})|\ y' = y\}| - |\{\boldsymbol{x}' \in \mathcal{K}_0(\boldsymbol{x})|\ y' = y\}|\Big).$$

If $y = 0$ and $\mathrm{diff}(\boldsymbol{x}) = t > 0$, then the negative label is more frequent among the neighbors of $\boldsymbol{x}$ within the protected group by a percentage of *t*. This suggests *t -discrimination* for the instance $(\boldsymbol{x}, y, a) \in D_1$.

Apart from flipping the value of *Y* or *A*, one can also perturb the distribution of the non-sensitive variables $\boldsymbol{X}$. Feldman et al. (2015) proposed an algorithm designed to learn an unbiased set $\{(\boldsymbol{x}_i, y_i, a_i)\}$ from any given dataset $\{(\boldsymbol{x}_i, y_i, a_i)\}$. For each instance, the algorithm updates only the value of $\boldsymbol{X}$, replacing it with $\boldsymbol{X}$ while keeping *Y* and *A* unchanged to preserve the utility of the classifier. Specifically, the algorithm uses the conditional probability $P(\boldsymbol{X} \leq \boldsymbol{x}_i \mid A = a_i)$ (assuming for simplicity that here $\boldsymbol{X} \in \mathbb{R}$) as the ranking for the tuple $(\boldsymbol{x}_i, y_i, a_i)$. It then substitutes $\boldsymbol{x}_i$ with an appropriately repaired value $\boldsymbol{x}_i$ that preserves this ranking. The authors propose different strategies for performing this substitution, which can be interpreted as a linear interpolation either in the ranking space between $\boldsymbol{X}$ and $\boldsymbol{X}$ or in the original value space of $\boldsymbol{X}$.

3.1.2 Learning Fair Representations

There exist methods that aim to find a discrimination-free representation set $\tilde{D} \subseteq \mathcal{X} \times \mathcal{Y} \times \{0,1\}$ for a given training set $D \subseteq \mathcal{X} \times \{0,1\} \times \{0,1\}$ and train the classifier on these representations instead. Such an approach differs from those introduced in Section 3.1.1 in two key ways. First, some methods (but not all of them) learn representations in a different value space, i.e., $\mathcal{X} \times \mathcal{Y} \neq \mathcal{X} \times \{0,1\}$ (Zemel et al., 2013; Madras et al., 2018), whereas fairness-enhancing methods that modify the data distribution always produce a new dataset valued in the same space as the original data. Second, the learning process of fair representations typically follows an adversarial framework. In this framework, one component seeks to minimize the information loss in $\tilde{D}$ relative to *D*, while the other aims to eliminate discrimination embedded in *D* from $\tilde{D}$.

Zemel et al. (2013) proposed an algorithm within this framework, which inspired a series of subsequent studies later on (Calmon et al., 2017; Lahoti et al., 2019; Beutel et al., 2017). They described this approach as “finding an intermediate representation of the data that best encodes the data.” Specifically, the ideal representations should simultaneously satisfy two objectives: (1) retaining as much information as possible and (2) excluding any information related to sensitive attributes. Zemel et al. (2013) introduced this approach as *Learned Fair Representations* (LFR). They formulated LFR as an optimization problem, designing the objective function to ensure group fairness, individual fairness, and accuracy.

To formally define LFR, let $Z \in \{1, \ldots, K\}$ denote the index of the *prototype*, where the prototype corresponding to $Z = k$ is associated with a vector $\boldsymbol{v}_k \in \mathbb{R}^d$, for all $k = 1, \ldots K$. Notably, the vectors $\{\boldsymbol{v}_k\}_{k=1}^{K}$ reside in the same space as $\boldsymbol{X}$. The objective is to learn an appropriate prototype *Z* such that: (1) the mapping from $\boldsymbol{X}$ to *Z* satisfies the desired fairness metrics; (2) *Z*, along with the associated vectors $\{\boldsymbol{v}_k\}_{k=1}^{K}$, retains as much information from $\boldsymbol{X}$ as possible; (3) the induced mapping from $\boldsymbol{X}$ to $\hat{Y}$ (through *Z*) closely approximates the ground-truth label *Y*.

Let $\mathrm{dist}(\cdot,\cdot)$ be a distance metric on $\mathbb{R}^d \times \mathbb{R}^d$. A common choice is the Euclidean distance, defined as $\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{v}_k) = \| \boldsymbol{x}_i - \boldsymbol{v}_k \|_2$. Then, a natural probabilistic mapping can be induced as

$$P(Z=k|\ \boldsymbol{X}=\boldsymbol{x})=\exp\{-\mathrm{dist}(\boldsymbol{x},\boldsymbol{v}_k)\}/\sum_{j=1}^{K}\exp\{-\mathrm{dist}(\boldsymbol{x},\boldsymbol{v}_j)\}.\ (27)$$

First, to ensure that the mapping from $\boldsymbol{X}$ to *Z* satisfies Demographic Parity, we require

$$P(Z=k|\ A=0)=P(Z=k|\ A=1),\quad \forall k=1,\ldots K.\ (28)$$

Define $M_{i,k}:=P(Z=k|\ X=\boldsymbol{x}_i)$ for all *i*, *k*. Here, $M_{i,k}$ represents the probability that the *i*th instance in the training set is assigned to the *k*th prototype. Recall that D_0 and D_1 denote the unprotected and protected groups, respectively. Then, Equation (28) can be approximated using the training data as

$$\underbrace{\frac{1}{|D_1|}\sum_{i:(\boldsymbol{x}_i,y_i,a_i)\in D_1}M_{i,k}}_{M_k^1}=\underbrace{\frac{1}{|D_0|}\sum_{i:(\boldsymbol{x}_i,y_i,a_i)\in D_0}M_{i,k}}_{M_k^0},\quad \forall k=1,\ldots,K.\ (29)$$

This leads to the loss function for enforcing Demographic Parity:

$$L_{\mathrm{parity}}=\sum_{k=1}^{K}|M_k^1-M_k^0|.\ (30)$$

Second, to ensure that *Z* serves as an informative representation of $\boldsymbol{X}$, we can reconstruct $\boldsymbol{X}$ from *Z* and minimize the difference between the original and reconstructed features. For each $\boldsymbol{x}_i$ from the given set *D*, a natural reconstruction approach is to define

$$\boldsymbol{x}_i:=\sum_{k=1}^{K}M_{i,k}\boldsymbol{v}_k,\quad \forall i=1,\ldots,N\ \ (31)$$

which represents a weighted combination of the prototype vectors. This induces the loss function controlling the information loss in the representation:

$$L_{\mathrm{descriptn}}=\sum_{i=1}^{N}||\boldsymbol{x}_i-\boldsymbol{x}_i\,||^2.\ (32)$$

Finally, to ensure the utility of predicting *Y* from *Z*, we define $w_k\in[0,1]$ as each prototype's prediction for *Y*, i.e., the estimated probability of a positive label given that an instance is assigned to prototype *k*. Then, the score that an instance with feature vector $\boldsymbol{x}_i$ relates to a positive label can be written as

$$s_i=\sum_{k=1}^{K}M_{i,k}w_k.\ (33)$$

This leads to the cross-entropy loss function for classification accuracy:

$$L_{\text{acc}} = -\sum_{i=1}^{N}\left[y_i \log s_i + (1-y_i)\log(1-s_i)\right]. \; (34)$$

With all the loss functions defined above, LFR can be written as the following optimization problem:

$$\min_{\boldsymbol{v}_k \in \mathbb{R}^d, w_k \in (0,1)} C_1 L_{\text{parity}} + C_2 L_{\text{descriptn}} + C_3 L_{\text{acc}} \; (35)$$

where $C_i, i = 1, 2, 3$ are the given constants specifying a balance between fairness, reconstruction, and accuracy. Notably, Problem (35) solves the fair representations (*Z* and the associated $\{\boldsymbol{v}_k\}$) as well as a classifier that predicts *Y* from *Z* based on $\{w_k\}$. However, the representations can be further applied to any downstream classification model independent of the parameters $\{w_k\}$. Therefore, this survey categorizes LFR as a pre-processing method.

LFR applies explicit regularization terms to ensure group fairness in the mapping from $\boldsymbol{X}$ to *Z*. Notably, this constraint also enforces Demographic Parity in the predictions, as Equation (29) directly implies $P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$. Similarly, while there is no explicit control over individual fairness for the predictions, the mapping mechanism ensures that instances with similar features ($\boldsymbol{X}$) are likely to be assigned to the same prototype and thus receive the same predicted label.

There are modified methods that follow Zemel et al. (2013)'s framework but elaborate on more diverse designs on the fair representations (Calmon et al., 2017; Madras et al., 2018; Lahoti et al., 2019). Calmon et al. (2017) learned representations for both $\boldsymbol{X}$ and *Y* through a pre-processing optimization problem that aims to minimize the distribution dissimilarity between $(\boldsymbol{X}, Y)$ and $(\boldsymbol{X}, \tilde{Y})$, subject to both group and individual fairness constraints. They kept the representation $(\boldsymbol{X}, \tilde{Y})$ within the same value space as the original data $(\boldsymbol{X}, Y)$. The group fairness constraint is formulated as

$$\left| \frac{P(\tilde{Y} = y | A = a_1)}{P(\tilde{Y} = y | A = a_2)} - 1 \right| < \epsilon_{a_1, a_2, y}, \quad \forall a_1, a_2 \in \{0,1\}, y \in \{0,1\} \; (36)$$

and the individual fairness constraint is written as

$$\mathbb{E}[\delta((\boldsymbol{x}, y), (\boldsymbol{X}, \tilde{Y})) | A = a, \boldsymbol{X} = \boldsymbol{x}, Y = y] < \sigma_{a,x,y}, \quad \forall \boldsymbol{x} \in \mathcal{X}, y, a \in \{0,1\} \; (37)$$

where $\delta : (\mathcal{X}, \{0,1\}) \times (\mathcal{X}, \{0,1\}) \to \mathbb{R}_+$ is a distortion metric with $\delta((\boldsymbol{x}, y), (\boldsymbol{x}, y)) = 0$, and ϵ and σ are non-negative thresholds. Similarly, Lahoti et al. (2019) proposed a method focused on eliminating individual unfairness from the representations. They learned representations via low-rank prototypes associated with vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$, $\boldsymbol{v}_k \in \mathbb{R}^d$ and $K < N$, as Zemel et al. (2013) did. However, instead of using the prototypes, this work treated the reconstruction $\boldsymbol{x}_i$ as the representation of $\boldsymbol{x}_i$. Meanwhile, they only included terms measuring reconstruction loss and individual fairness in the loss function, forming an unsupervised framework for performing a probabilistic transformation from the given data to individually fair representations.

3.1.3 Decoupling

Decoupling refers to techniques that train a separate classifier for each subgroup of the training set based on sensitive variables (Calders and Verwer, 2010; Dwork et al., 2018; Ustun et al., 2019). The most naive decoupling strategy is simply training one classifier for the protected group and another classifier for the unprotected group (Calders and Verwer, 2010). Decoupling can be seen as an additional pre-processing step before a black-box classifier. Unlike the methods in Section 3.1.1 and 3.1.2, which aim to eliminate information related to sensitive variables, the motivation behind decoupling is to appropriately leverage the sensitive attribute, as long as it is legal and ethical (Dwork et al., 2018).

Dwork et al. (2018) proposed a basic framework for decoupling that ensures desired parity-based metrics. Suppose that $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \ldots \cup \mathcal{X}_K$ is a partition of the value space $\mathcal{X}$ and $\pi : \mathcal{X} \to \{1, \ldots, K\}$ is an indicator function such that $\boldsymbol{x}_i \in \mathcal{X}_{\pi(\boldsymbol{x}_i)}$. For a set of *K* binary classifiers $\{f_1, \ldots, f_K\}$, where $f_k : \mathcal{X}_k \to [0,1]$. The decoupled classifier system $\gamma_f : \mathcal{X} \to [0,1]$ is defined as

$$\gamma_f(\boldsymbol{x}) = f_{\pi(\boldsymbol{x})}(\boldsymbol{x}). \quad (38)$$

Later, $\gamma_f(\boldsymbol{x})$ will be trained to minimize a joint loss function over all the separate classifiers f_k, $k = 1, \ldots, K$. By designing an appropriate joint loss function, one can enforce the desired fairness metrics. For example, Demographic Parity can be enforced by defining the loss function as

$$L_{\text{DP}} = \frac{\lambda}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| + (1-\lambda)\sum_{k=1}^{K}\left|\hat{p}_k \frac{N}{N_k} - \frac{1}{K}\sum_{k'=1}^{K}\hat{p}_{k'}\frac{N}{N_{k'}}\right| \quad (39)$$

where $N_k = \sum_{i=1}^{N}\mathbb{I}\{\pi(\boldsymbol{x}_i) = k\}$ represents the number of instances within each subgroup, and $\hat{p}_k = \frac{1}{N}\sum_{i:\pi(\boldsymbol{x}_i)=k}\hat{y}_i$ denotes the ratio of positive predictions within $\mathcal{X}_k$ relative to the total number of instances. The function (39) balances overall accuracy with the discrepancy in positive prediction ratios across subgroups.

Ustun et al. (2019) modified the Dwork et al. (2018)'s method to introduce a more dedicated partitioning strategy. Instead of considering the most granular groups based on all sensitive attributes for parity-based fairness metrics, they proposed a recursive searching algorithm for decoupling that satisfies (1) *rationality* and (2) *envy-freeness*. Suppose that *f* is a pooled classifier trained on a given dataset *D*, and let $D_k = \{(\boldsymbol{x}_i, y_i, a_i) \in D | \ \pi(\boldsymbol{x}_i) = k\}$, $k = 1, \ldots, K$ be the corresponding partitioning of the training set *D*. Rationality refers to the property that for each group *k*, the classifier f_k performs better (i.e., achieves higher accuracy) than *f* on D_k. Envy-freeness means that for each group *k*, f_k performs better than all other separate classifiers $f_{k'}$, for $k' \neq k$. This condition can be interpreted as ensuring that each group prefers its own classifier over others. The recursive searching algorithm resembles the decision tree algorithm but is constrained by both rationality and envy-freeness. This decoupling approach allows training the most accurate model for each group without negatively impacting others.

3.2 In-processing

In general, in-processing fairness-enhancing mechanisms are applied during the training phase of machine learning algorithms. These methods typically involve directly modifying model parameters, architectures, or training procedures. Therefore, they often require users to have control over the target model and the training process.

For clarity regarding the effect of model parameters, this section explicitly writes $f(\boldsymbol{x}, \boldsymbol{\theta})$ as the model output for features $\boldsymbol{x} \in \mathcal{X}$ and parameters $\boldsymbol{\theta} \in \Theta$.

3.2.1 Regularizer and Fairness Constraints

In general, a machine learning model is trained by solving an optimization problem where the loss function $L(\boldsymbol{\theta}, D)$ serves as the objective function, and the parameters $\boldsymbol{\theta}$ serve as the decision variables. While such an optimization problem is typically designed to maximize accuracy, it can be modified to enforce fairness by either (1) adding a fairness-related regularization or penalty term to the loss function or (2) incorporating additional fairness-enhancing constraints.

Kamishima et al. (2012) proposed a prejudice-removal regularizer to reduce indirect discrimination in binary classifiers. They assumed that the machine learning model embeds no direct discrimination, i.e., the sensitive variable is not used in the classifier, while $\hat{Y}$ and A may still be statistically dependent, implying indirect discrimination. To quantify this dependence, the authors introduced the *indirect Prejudice Index* (PI), which measures the mutual information between a target variable Y and the sensitive attribute A:

$$\mathrm{PI} = \sum_{(y,a)\in\{0,1\}\times\mathcal{A}} P(Y=y, A=a) \ln \frac{P(Y=y, A=a)}{P(Y=y)P(A=a)}.$$

Let $D = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^{N}$ denote the training set and $f(\cdot, \boldsymbol{\theta})$ the classifier that does not access A. After omitting scaling factors, the PI for f can be approximated as

$$\mathrm{PI} = \sum_{i=1}^{N} \sum_{y\in\{0,1\}} \hat{P}_f(Y=y \mid \boldsymbol{X}=\boldsymbol{x}_i) \ln \frac{\hat{P}_f(Y=y \mid A=a_i)}{\hat{P}_f(Y=y)}, \quad (40)$$

where $\hat{P}_f(\cdot)$ denotes the probabilities induced by the output of f. Specifically, $\hat{P}_f(Y=y \mid \boldsymbol{X}=\boldsymbol{x}_i) = yf(\boldsymbol{x}_i, \boldsymbol{\theta}) + (1-y)(1-f(\boldsymbol{x}_i, \boldsymbol{\theta}))$ for $y \in \{0,1\}$. The quantities $\hat{P}_f(Y=y \mid A=a)$ and $\hat{P}_f(Y=y)$ are computed as

$$\hat{P}_f(Y=y\mid A=a)=\frac{\sum_{i:a_i=a}\hat{P}_f(Y=y\mid \boldsymbol{X}=\boldsymbol{x}_i)}{\sum_{i=1}^{N}\mathbb{I}(a_i=a)},$$

$$\hat{P}_f(Y=y)=\frac{\sum_{i=1}^{N}\hat{P}_f(Y=y\mid \boldsymbol{X}=\boldsymbol{x}_i)}{N}.$$

Equation (40) can thus be fully determined by the model $f(\cdot,\boldsymbol{\theta})$ and the training data *D*. It can be incorporated as a regularization term in the objective function to promote fairness during model training.

Beyond designing regularizers, another common approach is to introduce fairness-enhancing constraints into the optimization problem. For example, to prevent Disparate Treatment, one can impose the following constraints (Zafar et al., 2017a):

$$\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta},D) \quad \text{s.t. } |P(\hat{Y}\neq Y\mid A=0)-P(\hat{Y}\neq Y\mid A=1)|<\epsilon \quad (41)$$

where *L* is the loss function and ϵ is a positive constant. Notably, as Problem (41) suggests, constraints directly derived from fairness metrics are often non-convex. The key challenge lies in formulating constraints that enforce the desired fairness property while remaining computationally tractable.

Zafar et al. (2017b) proposed a formulation that ensures *p*%-rule (Demographic Parity) based on the distance from the feature vectors to the decision boundary. Let $\{\text{dist}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\}_{i=1}^{N}$ be the signed distance from features $\boldsymbol{x}_i, i=1,\dots,N$ to the decision boundary of the classifier parameterized by $\boldsymbol{\theta}$. $\text{dist}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\geq 0$ implies $\hat{Y}=1$ and $\text{dist}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)<0$ implies $\hat{Y}=0$. The covariance between the sensitive attributes *A* and the distances can be approximated by

$$\text{Cov}(A,\text{dist}_{\boldsymbol{\theta}}(\boldsymbol{X}))\approx\frac{1}{N}\sum_{i=1}^{N}(a_i-\bar{a})\text{dist}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \quad (42)$$

where $\bar{a}=\frac{1}{N}\sum_{i=1}^{N}a_i$. Expression (42) is a convex function with respect to $\boldsymbol{\theta}$ for all linear, convex margin-based classifiers. The optimization problem can then be formulated as

$$\begin{aligned}&\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta},D)\\ &\text{s.t. } \left|\frac{1}{N}\sum_{i=1}^{N}(a_i-\bar{a})\ \text{dist}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right|<c\end{aligned} \quad (43)$$

where *c* is a positive constant that balances accuracy and fairness. As $c\to 0$, the resulting classifier satisfies a stricter *p*%-rule but is constrained to a smaller set of feasible parameters, potentially leading to higher accuracy loss. Extensions of the optimization problem (43) to other group fairness notions can be found in Zafar et al. (2017a)'s work.

It is well known that regularization terms and constraints are interchangeable in an optimization problem. Therefore, this survey does not distinguish between these two strategies in the subsequent discussions. More recent methods have introduced modifications to enhance interpretability and generalizability. For instance, Ross et al. (2017) integrated gradient-based interpretation with fairness control, designing regularization terms that suppress irrelevant gradients, ensuring that the classifier makes decisions for the right reasons. Agarwal et al. (2018) proposed group fairness constraints for Demographic Parity and Equalized Odds, which can be applied to randomized model sampling from a distribution over the space of all possible classifiers. Celis et al. (2019) further refined the optimization framework, demonstrating that non-convex linear fractional fairness constraints can be effectively handled.

The discussion in this section has focused on the optimization framework for binary classifiers that enforce commonly used fairness notions. However, the design of regularization terms and constraints can be readily extended to a wide range of machine learning models. For instance, Berk et al. (2017) proposed an optimization framework for regression models that satisfy both group and individual fairness metrics. Meanwhile, some methods enforce fairness through additional rules embedded in the training algorithms rather than explicit regularization terms or constraints. For example, Roh et al. (2020) developed a batch selection algorithm called FairBatch, which adjusts batch sizes for the protected and unprotected groups based on fairness metrics measured in the current epoch.

3.2.2 Fair Adversarial Learning

Fair adversarial learning (Beutel et al., 2017) refers to methods that seek to obtain a fair intermediate representation through adversarial training (Goodfellow et al., 2014; Bousmalis et al., 2016; Ganin et al., 2016). This approach follows a similar concept to LFR, as discussed in Section 3.1.2. We classify these methods as in-processing techniques, as their implementation typically involves modifying the model structure.

Edwards and Storkey (2016) proposed a general loss function for the task of finding intermediate fair representations $\boldsymbol{Z} \in \mathbb{R}^m$ for inputs $\boldsymbol{X} \in \mathbb{R}^d$, which can be expressed as

$$L(\boldsymbol{\theta},\phi) = C_1 L_{\text{parity}}(A,\boldsymbol{Z},\boldsymbol{\theta},\phi) + C_2 L_{\text{reconst}}(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}) + C_3 L_{\text{acc}}(\boldsymbol{Z},Y,\boldsymbol{\theta}) \quad (44)$$

where the optimization problem is formulated as the following minimax function:

$$\min_{\boldsymbol{\theta}} \max_{\phi} L(\boldsymbol{\theta},\phi) \quad (45)$$

with $C_i > 0$ for $i = 1,2,3$. Here, L_{reconst} represents the cost of reconstructing $\boldsymbol{X}$ from $\boldsymbol{Z}$, L_{acc} denotes the error in predicting *Y* from $\boldsymbol{Z}$, and L_{parity} measures the dependence between $\boldsymbol{Z}$ and *A*. In general, L_{parity} can be defined as the loss function of a classifier that distinguishes between the distributions $P(\boldsymbol{Z} \mid A = 0)$ and $P(\boldsymbol{Z} \mid A = 1)$. Such a classifier, parameterized by ϕ, is referred to as an *adversary*. Since $\boldsymbol{Z}$ should retain as little information about *A* as possible, the optimization problem in (45) first maximizes the loss with respect to the adversary, thereby enhancing fairness, and then minimizes the loss associated with prediction and reconstruction.

Learning fair representations involves multiple competing objectives. Beutel et al. (2017) applied adversarial training to remove information related to sensitive attributes from intermediate representations. They designed a multi-head deep neural network, where one head predicts the target label *Y*, while the other is adversarially trained to prevent the prediction of the sensitive attribute *A*. Specifically, let $f:\mathcal{X}\to\mathcal{Z}$ be the mapping from $\boldsymbol{X}$ to the intermediate representation $\boldsymbol{Z}$, $g:\mathcal{Z}\to[0,1]$ be the classifier predicting *Y* from $\boldsymbol{Z}$, and $h:\mathcal{Z}\to\mathcal{A}$ be the classifier predicting *A* from $\boldsymbol{Z}$. The functions *f*, *g*, and *h* can be arbitrary neural networks and are trained simultaneously. The multi-head model is optimized using the following objective:

$$\min\sum_{i=1}^{N} L_Y(g(f(\boldsymbol{x}_i)),y_i)+L_A(h(J_\lambda(f(\boldsymbol{x}_i))),a_i) \quad (46)$$

where L_Y and L_A are the loss functions measuring the prediction errors for *Y* and *A*, respectively. The function J_λ is an identity transformation with a negative gradient:

$$J_\lambda(f(\boldsymbol{x}))=f(\boldsymbol{x}),\quad \frac{dJ_\lambda}{d\boldsymbol{x}}=-\lambda\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}$$

where λ is a positive constant that controls the trade-off between accuracy and fairness. The optimization problem in (46) encourages the prediction of *Y* from $\boldsymbol{X}$ through $\boldsymbol{Z}$, while simultaneously discouraging the prediction of *A* from $\boldsymbol{X}$ through $\boldsymbol{Z}$.

Madras et al. (2018) proposed a general framework for fair adversarial learning, consisting of an encoder mapping $\boldsymbol{X}$ to $\boldsymbol{Z}$, a decoder mapping $\boldsymbol{Z}$ back to $\boldsymbol{X}$, a classifier predicting *Y* from $\boldsymbol{Z}$, and an adversary (classifier) predicting *A* from $\boldsymbol{Z}$, as illustrated in Figure 1. The model is typically trained to minimize the loss associated with the decoder and the classifier while maximizing the adversary's loss. The adversarial loss function can be designed based on the chosen fairness metrics. In practice, the decoder can be omitted (Beutel et al., 2017), allowing the intermediate representation $\boldsymbol{Z}$ to be used for arbitrary downstream tasks.

Furthermore, Abusitta et al. (2020) integrated Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) with the original classifier to learn fair representations of the inputs. They trained a generator to synthesize data for the group that the classifier discriminated against and then used the generated data to augment the training set, enabling the training of a fairer model. On the other hand, Xu et al. (2018) applied GANs entirely at the pre-processing stage. They proposed FairGAN, which consists of one generator and two discriminators. The generator synthesizes training data, while one discriminator distinguishes between real and generated data, and the other discriminator determines whether a generated instance belongs to the protected or unprotected group.

Similar to the methods discussed in Section 3.1.2, the approaches introduced in this section can often be applied in either an in-processing or pre-processing manner, as the fair synthesized data generated through adversarial learning can be seamlessly integrated into training arbitrary downstream machine learning models.

3.3 Post-processing

Post-processing techniques are methods designed to enforce fairness after the training process is complete. These techniques do not require modifications to the training process or the model itself but instead operate on the model's predictions or inferred outcomes.

One post-processing strategy involves perturbing the original predictions $\hat{Y}$ to produce unbiased results $\tilde{Y}$. A common approach is thresholding, which assigns customized thresholds for output scores based on group membership (Corbett-Davies et al., 2017):

$$\tilde{y}_i = \begin{cases} 1 & f(\boldsymbol{x}_i) \geq c_{a_i} \\ 0 & \text{otherwise} \end{cases}$$

where c_{a_i} is a group-specific threshold that varies based on the group membership indicated by the sensitive attribute. The appropriate values of c_{a_i} are chosen to ensure that the outcomes satisfy the specified (group) fairness metrics.

Thresholding can be viewed as a special case of learning a predictor $\tilde{Y}$ derived from the original output *R* and the sensitive attribute *A*, ensuring that $\tilde{Y}$ is independent of $\boldsymbol{X}$ given (R, A) (Hardt et al., 2016). Here, *R* is a random variable representing the model output, which can take different forms. For example, it may correspond to the predicted label $\hat{Y}$, the estimated conditional probability $\hat{P}(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})$, or a numeric score used for decision-making. Hardt et al. (2016) provided the following example for deriving $\tilde{Y}$ when $R = \hat{Y} \in \{0,1\}$. Define $p_{y,a} := P(\tilde{Y}=1 \mid \hat{Y}=1, A=a)$, let $p := (p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1})$, and define $\gamma_a(\tilde{Y}) := (P(\tilde{Y}=1 \mid Y=0, A=a), P(\tilde{Y}=1 \mid Y=1, A=a))$. Let $\tilde{Y}_p$ be the derived predictor characterized by *p*. Then the optimization problem ensures that $\tilde{Y}$ satisfies Equalized Odds is formulated as

$$\begin{aligned} &\min_p \mathbb{E}[L(\tilde{Y}_p, Y)] \\ &\text{s.t.} \quad \gamma_0(\tilde{Y}_p) = \gamma_1(\tilde{Y}_p) \qquad (47) \\ &0 \leq p_{y,a} \leq 1, \quad \forall y, a \in \{0,1\} \end{aligned}$$

where $L: \{0,1\}^2 \rightarrow \mathbb{R}$ is the given loss function for prediction. This optimization problem aims to minimize the prediction loss of $\tilde{Y}$ given the ground-truth *Y*, subject to the constraints directly taken from Equation (4).

Woodworth et al. (2017) combined in-processing and post-processing strategies to propose a two-step fairness-enhancing algorithm. In the first step, an approximately non-discriminatory predictor $\hat{Y}$ is estimated by incorporating fairness constraints. In the second step, a randomized predictor is derived from $(\hat{Y}, Y, A)$ to $(\tilde{Y}, Y, A)$ to further reduce discrimination. On the other hand, Jiang et al. (2020) introduced a technique that directly perturbs $f(\boldsymbol{x}_i)$ to enforce independence between the output and the sensitive variables. The perturbation is applied to minimize Wasserstein-1 distances, ensuring that the perturbed data follows the desired, discrimination-free distribution.

Another strategy involves perturbing the input $\boldsymbol{X}$ using a pre-processor designed to align with a well-trained classifier. Wang et al. (2019) introduced a pre-processor *T* that can be applied to any black-box classifier *f*. The resulting classifier $\tilde{f}$ is defined as follows:

$$\tilde{f}(\boldsymbol{X}) = \begin{cases} f(T(\boldsymbol{X})) & \text{if } A = 1, \\ f(\boldsymbol{X}) & \text{otherwise.} \end{cases}$$

Here, *T* is designed to mitigate discrimination with respect to specific (group) fairness metrics by perturbing the distribution of features from the protected group, $D_1 = \{(\boldsymbol{x}_i, y_i, a_i) \in D |\ a_i = 1\}$. The authors assumed the existence of a *counterfactual distribution*, which refers to a hypothetical probability distribution for the protected group such that the performance gap concerning the specified fairness metrics is minimized. They further demonstrated that this counterfactual distribution can be recovered by reweighting D_1, where the weights are computed based on repeated inference outcomes of the well-trained models, guided by the desired fairness metrics. It is worth noting that the pre-processor *T* is learned and applied after the training of *f*, operating solely when inferring the well-trained model. Therefore, this survey classifies this method as a post-processing approach. Similarly, Berk et al. (2024) proposed a novel framework improving the fairness of risk assessment tools. They suggested training classifiers on the more privileged group (unprotected group) and then transferring the distribution of a less privileged group (protected group) to the more privileged one for predictions. Such an approach helps achieve Predictive Parity in model predictions.

This paper has thus far summarized key fairness-enhancing mechanisms for binary classifiers. Notably, fairness is also gaining attention in more complex machine learning algorithms, such as regression (Calders et al., 2013; Chzhen and Schreuder, 2022), causal inference (Kusner et al., 2017), and recommendation systems (Yang and Stoyanovich, 2017), operating on diverse types of data, including texts (Bolukbasi et al., 2016), images (Karkkainen and Joo, 2021), and sequential data (Gonzalez et al., 2017). Due to space limitations, we will discuss these extensions in the Appendices.

4 Experiments

This section aims to give the reader a straightforward overview of the implementations of all three kinds of fairness-enhancing methods introduced in Section 3.

4.1 Experimental Setting

We conducted experiments on the Adult (Dua and Graff, 2019) and COMPAS (ProPublica, 2016) datasets across binary classifiers based on logistic regression, support vector machines (SVMs), and multilayer perceptrons (MLPs). The selected fairness-enhancing methods include the pre-processing approach LAFTR (Madras et al., 2018), the in-processing approaches LabelDebias (Jiang and Nachum, 2020) and FairConstr (Zafar et al., 2017b), and a naive post-processing approach, Threshold, implemented by the authors. This section leverages five metrics to assess each method’s effectiveness: (1) Acc, the overall accuracy of the predictions; (2) AccProt, the accuracy of predictions within the protected group; (3) Disparity, the discrepancy in positive decisions between the protected and unprotected group; (4) Opty, the difference in True Positive Rates between the protected and unprotected group; and (5) Odds, the difference in False Positive Rates between the protected and unprotected groups. Acc and AccProt range from *[*0, 1*]*, with higher values indicating greater classifier utility. The fairness metrics Disparity, Opty, and Odds range from $(-1,1)$, where values closer to 0 indicate greater fairness. Positive values for these fairness metrics suggest potential discrimination against the protected group in model outputs. For further details on the experimental settings, please refer to the Appendices.

4.2 Results

According to the results listed in Table 3, we conclude that machine learning models tend to inherit or even magnify the discrimination embedded in the input data, and the fairness-enhancing methods are capable of mitigating such a magnification. However, there are two trade-offs worth noticing. First, there is a trade-off between model utility and fairness, as a decrease in Disparity (i.e., increase in fairness level) often leads to reduced prediction accuracy. Second, there is a trade-off between different fairness metrics. For example, a decrease in Disparity may lead to an increase in Opty and Odds. Despite the incompatibility we discussed in Section 2.3, experimental results suggest that trade-offs may also arise among other combinations of fairness notions.

It is important to keep in mind that, beyond the desired fairness metrics and acceptable trade-offs, the choice of fairness-enhancing methods also depends on the available information and resources. As discussed in Section 3, pre-processing methods require access to both training and testing data, in-processing methods require control over the training algorithms, model structures, and training process, while post-processing methods require the ability to infer from well-trained models. Additionally, the computational cost is an essential consideration. For instance, while LabelDebias achieves high fairness levels while maintaining high accuracy, it requires repeated model training and weight updates based on each iteration’s output. This process can be resource-intensive, particularly for complex models or large datasets.

5 Discussion

This paper presents a guided overview of fairness-enhancing mechanisms, along with their target fairness criteria and experimental evaluations of representative techniques. We organize key methods according to the machine learning lifecycle, categorizing pre-, in-, and post-processing approaches based on the required level of access and control over the model and data. Additionally, we highlight the specific fairness metrics each method aims to improve and discuss their inherent trade-offs, providing practical guidance for selecting appropriate techniques based on available resources and fairness goals.

Despite the growing number of studies in this area, several challenges hinder the broader development and adoption of machine learning fairness. First, compared to group fairness, individual fairness has received relatively little focus. This may be due to the task-dependent nature of individual fairness metrics, which makes them harder to define and compare across applications. Moreover, their computational cost arises from the need to assess each input-output pair, which can be prohibitive for large-scale datasets. Second, the field suffers from inconsistent use of terminology. Multiple names may be assigned to similar or the same fairness metrics. Such inconsistency can obscure understanding and hinder communication across studies.

In conclusion, as machine learning systems become increasingly integrated into high-stakes decision-making processes, it is crucial to develop models that are not only accurate but also fair and transparent. Addressing fairness concerns is essential to building trustworthy systems that equitably serve all members of society.

6 Acknowledgements

Funding: This work was supported in part by the National Science Foundation under grants DMS-2152289, and DMS-2134107.

AI policy confirmation: The authors used ChatGPT (version 4.5) for language polishing, LaTeX formatting, and debugging assistance, with all modifications thoroughly verified by the authors. All content, ideas, and algorithmic logic remain entirely the authors' own.

DAS: The datasets used in this study are publicly available. The UCI Adult dataset can be accessed at https://archive.ics.uci.edu/dataset/2/adult. The COMPAS dataset is available at https://github.com/propublica/compas-analysis.

Reproducibility: The codes and experimental results can be accessed at https://github.com/Xianwen-He/Fair_Bench.git.

7 Conflict of Interest

The authors report there are no competing interests to declare.

References

Abusitta, A., Aïmeur, E., and Abdel Wahab, O. (2020). Generative adversarial networks for mitigating biases in machine learning systems. In *ECAI 2020*, pages 937–944. IOS Press.

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Agarwal, A., Dudík, M., and Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: Risk assessments in criminal sentencing. *ProPublica*.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021a). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021b). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Berk, R. A., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2023). Fair risk algorithms. *Annual Review of Statistics and Its Application*, 10(1):165–187.

Berk, R. A., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024). Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *Sociological Methods & Research*, 53(4):1629–1675.

Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., and Chi, E. H. (2019). Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Biddle, D. (2017). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). Domain separation networks. *Advances in neural information processing systems*, 29.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Burke, R., Sonboli, N., and Ordonez-Gauger, A. (2018). Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on fairness, accountability and transparency*, pages 202–214. PMLR.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pages 71–80. IEEE.

Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328.

Celis, L. E. and Keswani, V. (2019). Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443*.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., and Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chouldechova, A. and G’Sell, M. (2017). Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*.

Chouldechova, A. and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Chzhen, E. and Schreuder, N. (2022). A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’17, pages 797–806. Association for Computing Machinery.

Cotter, A., Jiang, H., Gupta, M., Wang, S., Narayan, T., You, S., and Sridharan, K. (2019). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4):1–36.

Du, M., Yang, F., Zou, N., and Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34.

Dua, D. and Graff, C. (2019). UCI machine learning repository: Adult data set.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133. PMLR.

Edwards, H. and Storkey, A. (2016). Censoring representations with an adversary. In *4th International Conference on Learning Representations*, pages 1–14.

Eloundou, T., Beutel, A., Robinson, D. G., Gu-Lemberg, K., Brakman, A.-L., Mishkin, P., Shah, M., Heidecke, J., Weng, L., and Kalai, A. T. (2024). First-person fairness in chatbots. *arXiv preprint arXiv:2410.19803*.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. (2022). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152.

Faliagka, E., Ramantas, K., Tsakalidis, A., and Tzimas, G. (2012). Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*, pages 215–220.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

FiveThirtyEight (2015). Should prison sentences be based on crimes that haven't been committed yet? Accessed: 2025-03-24.

Flores, A. W., Bechtel, K., and Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". *Federal Probation*, 80(2).

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Gao, B., Wang, Y., Xie, H., Hu, Y., and Hu, Y. (2023). Artificial intelligence in advertising: advancements, challenges, and ethical considerations in targeting, personalization, content creation, and ad optimization. *Sage Open*, 13(4):21582440231210759.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. *Advances in neural information processing systems*, 29.

Gonzalez, C., Fakhari, P., and Busemeyer, J. (2017). Dynamic decision making: Learning processes and new research directions. *Human factors*, 59(5):713–721.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *Symposium on Machine Learning and the Law*.

Hajian, S. and Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

IBM Watson Services (2017). Watson Visual Recognition. Accessed: 2017-10-06.

Jiang, H. and Nachum, O. (2020). Identifying and correcting label bias in machine learning. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 702–712. PMLR.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2020). Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR.

Jung, C., Kannan, S., Lee, C., Pai, M., Roth, A., and Vohra, R. (2020). Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 677–678.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, pages 924–929. IEEE.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In Flach, P. A., De Bie, T., and Cristianini, N., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg. Springer Berlin Heidelberg.

Karkkainen, K. and Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558.

Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Kovalchuk, O., Teremeckyi, V., Kolesnikov, A., Chudyk, N., and Kaniuka, V. (2024). Machine learning models for information support in the justice system. In *2024 14th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 762–765. IEEE.

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., and Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862.

Kučak, D., Juričić, V., and Džambić, G. (2018). Machine learning in education-a survey of current research trends. *Annals of DAAAM & Proceedings*, 29.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.

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

Liao, C., Todo, Y., Wang, M., and Pan, Y. (2024). Machine learning for talent analytics: Unveiling competency indicators in live streamer. In *2024 IEEE International Conference on Cognitive Computing and Complex Data (ICCD)*, pages 189–194. IEEE.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE.

Luan, H. and Tsai, C.-C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1):250–266.

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

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR.

Megvii Inc. (2017). Face++ API. Accessed: 2017-10-06.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Microsoft Cognitive Services (2017). Microsoft Face API. Accessed: 2017-10-06.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163.

Neuberg, L. G. (2003). Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685.

Northpointe Inc. (2010). Compas risk & need assessment system: Selected questions posed by inquiring agencies. https://www.scribd.com/document/389260049/FAQ-Document-pdf. Accessed: 2025-06-15.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568.

Pessach, D. and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., and Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision support systems*, 134:113290.

Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. (2021). Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955.

Pfeiffer, J., Gutschow, J., Haas, C., Möslein, F., Maspfuhl, O., Borgers, F., and Alpsancar, S. (2023). Algorithmic fairness in ai: an interdisciplinary view. *Business & Information Systems Engineering*, 65(2):209–222.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, 30.

ProPublica (2016). Compas risk assessment data and analysis.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. (2020). Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., and Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458.

Shekhar, S., Fields, G., Ghavamzadeh, M., and Javidi, T. (2021). Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34:24535–24544.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2239–2248.

Ustun, B., Liu, Y., and Parkes, D. (2019). Fairness without harm: Decoupled classifiers with preference guarantees. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6373–6382. PMLR.

VanderWeele, T. J. and Hernán, M. A. (2012). Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *American journal of epidemiology*, 175(12):1303–1310.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Wan, M., Zha, D., Liu, N., and Zou, N. (2023). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27.

Wang, H., Ustun, B., and Calmon, F. (2019). Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6618–6627. PMLR.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR.

Xu, D., Yuan, S., Zhang, L., and Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, pages 570–575. IEEE.

Yang, K. and Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

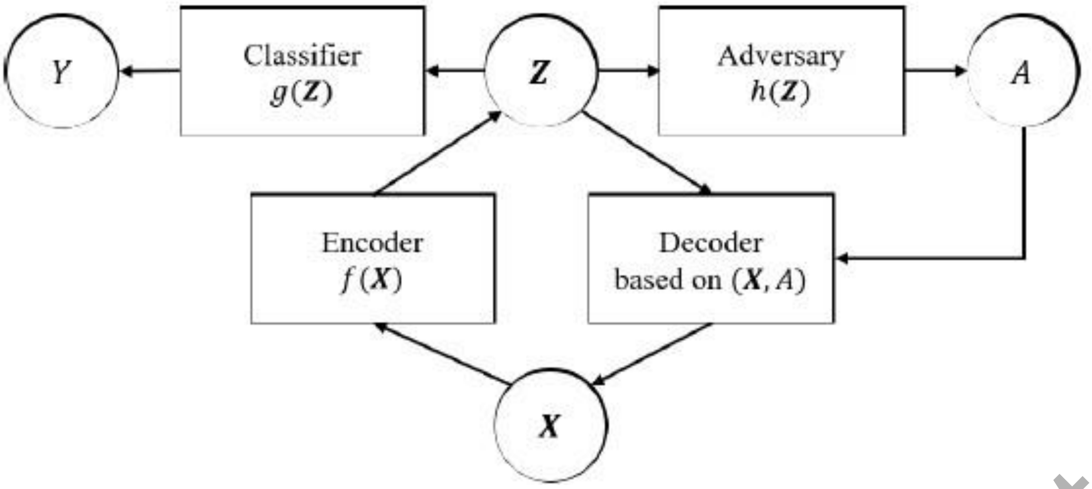

Figure 1: The framework for learning adversarial fair representations proposed by Madras et al. (2018). This figure is modified from Madras et al. (2018)'s work.

Table 1: Summary of fairness metrics.

Category	Metric	Feature
Group Fairness	Demographic Parity	The proportion of positive predictions is identical across protected and unprotected groups.
	Equalized Odds	True Positive Rates and False Positive Rates are identical across protected and unprotected groups.
	Equalized Opportunity	True Positive Rates are identical across protected and unprotected groups.
	Overall Accuracy Equality	Prediction accuracy is identical across protected and unprotected groups.
	Disparate Treatment	The sensitive attribute has no effect on the conditional distribution of predicted labels given the features.
	Conditional Equality of Opportunity	Given a specified feature, for every possible value of that feature, the False Positive Rates are identical across protected and unprotected groups.
	Predictive Parity	The predicted score accurately reflects the proportion of positive labels within each protected/unprotected group.
Individual Fairness	Consistency	Predictions for each instance are consistent with those of its neighbors in the feature space.
	Counterfactual Fairness	A model’s decision for an individual remains unchanged in a counterfactual world where the individual belongs to a different demographic group.
	Generalized Entropy Index	Prediction errors for individuals are close to the overall model performance.

Table 2: Overview of fairness-enhancing methods. The advantages and disadvantages in the table apply to the majority of methods under the corresponding category. DP: Demographic Parity. (*): The method can be extended to additional fairness metrics not listed in the original reference.

Pre-Processing						
Sub-Category	Feature	Advantages	Disadvantages	Representative Method	Applied Metrics	Related Work
Reweighting and Sampling	Assign customized weights to training instances.	Simple to implement and easy to interpret; flexible.	May distort the original data distribution and introduce extra noise.	naive reweighting (Kamiran and Calders, 2012)	DP	Kamiran and Calders (2012)
				label debias (Jiang and Nachum, 2020)	DP, Equal Opportunity, Equalized Odds (*)	Jiang and Nachum (2020)
Relabeling and Feature Perturbation	Modify values of labels, sensitive attributes, or features.	Same as above.	Same as above.	label flipping (Kamiran and Calders, 2012)	DP	Kamiran and Calders (2012)
				counterpart comparison (Luong et al., 2011)	individual fairness metric similar to consistency	Luong et al. (2011)
				feature perturbation (Feldman et al., 2015)	DP	Feldman et al. (2015)
Learning Fair Representation	Learn an intermediate representation that encodes inputs while excluding sensitive information.	Enables trade-offs among utility, fairness, and preserved information from the training set.	May distort the original information in the training set.	Learned Fair Representations (Zemel et al., 2013)	DP, consistency	Zemel et al. (2013) Calmon et al. (2017) Lahoti et al. (2019)
Decoupling	Train a separate classifier for each subgroup	Preserves the original distribution and data characteristics.	Requires training multiple classifiers.	decoupling (Dwork et al., 2018)	DP, Treatment Equality, Equalized Odds (*)	Dwork et al. (2018) Ustun et al. (2019)

	defined by sensitive attributes.					
In-Processing						
Regularizer and Fairness Constraints	Add regularization terms or fairness constraints to the optimization problem.	Straightforward and interpretable in optimization settings.	Difficult for non-convex problems or metrics; the training algorithm aware of one fairness metric is often difficult to apply to another metric directly.	prejudice-removal regularizer (Kamishima et al., 2012)	fairness metric similar to Disparate Treatment	Kamishima et al. (2012)
				fairness-enhancing constraints (Zafar et al., 2017b)	DP, Disparate Treatment	Zafar et al. (2017a) Zafar et al. (2017b)
				generalized fairness constraints	-	Ross et al. (2017) Agarwal et al. (2019) Celis and Keswani (2019)
Fair Adversarial Learning	Incorporate an adversary to reduce the influence of sensitive attributes.	Similar to Learning Fair Representation.	Increases training complexity and may require model modifications.	classifier-adversary learning (Edwards and Storkey, 2016)	DP	Edwards and Storkey (2016) Beutel et al. (2017) Madras et al. (2018)
				GAN-based learning (Abusitta et al., 2020)	DP	Abusitta et al. (2020) Xu et al. (2018)
Post-Processing						
	Modify predictions or inputs after training.	Works with black-box models.	May significantly harm predictive accuracy; sometimes requires	thresholding (Hardt et al., 2016)	DP, Disparate Treatment, Overall Accuracy Equality	Hardt et al. (2016)
				post-processor	Equalized	Woodwort

			repeated retraining.	based on outputs (Woodworth et al., 2017)	Odds	h et al. (2017) Jiang et al. (2020)
				pre-processor based on outputs (Wang et al., 2019)	DP, Equalized Odds, Equal Opportunity	Wang et al. (2019) Berk et al. (2024)

Table 3: Evaluation of four fairness-enhancing methods across three base models on the Adult and COMPAS datasets, compared to baseline model performances without fairness guarantees. All scores are reported as percentages. The implementation of FairConstr is limited to models with convex loss functions and does not apply to MLP. The first row of the table measures the potential discrimination embedded in the original dataset.

Model	Fair Method	Adult					COMPAS				
		Acc	AccProt	Disparity	Opty	Odds	Acc	AccProt	Disparity	Opty	Odds
	Raw Data	-	-	12.74	-	-	-	-	4.98	-	-
Logistic	baseline	85.84	92.73	11.72	8.57	4.25	67.69	67.50	11.36	9.55	9.71
	LAFTR	82.66	90.28	3.09	-15.40	-0.31	67.63	66.42	4.82	3.57	2.75
	LabelDebias	84.45	90.78	2.35	-19.83	-1.52	67.80	68.33	1.48	-1.02	0.48
	FairConstr	59.57	46.71	-6.38	-12.53	-10.99	54.93	49.95	0.00	0.00	0.00
	Threshold	85.02	91.03	5.14	-11.73	-0.01	67.80	67.77	6.74	4.61	5.41
SVM	baseline	75.16	85.10	15.27	10.83	9.95	65.52	64.17	10.09	8.96	8.34
	LAFTR	42.66	42.18	9.94	12.57	8.02	68.02	67.53	4.89	2.59	3.87
	LabelDebias	68.72	78.01	10.70	5.55	7.75	66.57	67.22	-0.25	-2.87	-0.86
	FairConstr	85.64	92.64	10.93	7.41	3.79	65.52	64.17	10.09	8.96	8.34
	Threshold	74.93	81.61	8.69	-0.86	4.34	66.24	65.19	5.30	3.68	3.92
MLP	baseline	82.78	90.19	11.85	5.66	5.23	64.18	62.60	9.21	9.29	6.33
	LAFTR	84.56	90.97	9.52	-0.07	3.00	65.68	65.22	6.42	4.46	5.48
	LabelDebias	83.24	90.41	11.08	6.65	4.44	65.63	64.36	10.85	10.51	8.11
	FairConstr	-	-	-	-	-	-	-	-	-	-
	Threshold	82.18	88.01	5.07	-13.14	0.46	64.01	63.80	5.08	3.76	3.63