Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification

Technical Report

Maliha Tashfia Islam, Anna Fariha, Alexandra Meliou
College of Information and Computer Sciences, University of Massachusetts, Amherst
{mtislam, afariha, ameli}@cs.umass.edu

ABSTRACT

Classification, a heavily-studied data-driven machine learning task, drives an increasing number of prediction systems involving critical human decisions such as loan approval and criminal risk assessment. However, classifiers often demonstrate discriminatory behavior, especially when presented with biased data. Consequently, fairness in classification has emerged as a high-priority research area. Data management research is showing an increasing presence and interest in topics related to data and algorithmic fairness, including the topic of fair classification. The interdisciplinary efforts in fair classification, with machine learning research having the largest presence, have resulted in a large number of fairness notions and a wide range of approaches that have not been systematically evaluated and compared. In this paper, we contribute a broad analysis of 13 fair classification approaches and additional variants, over their correctness, fairness, efficiency, scalability, and stability, using a variety of metrics and real-world datasets. Our analysis highlights novel insights on the impact of different metrics and high-level approach characteristics on different aspects of performance. We also discuss general principles for choosing approaches suitable for different practical settings, and identify areas where data-management-centric solutions are likely to have the most impact.

1 INTRODUCTION

Virtually every aspect of human activity relies on automated systems that use prediction models learned from data: from routine everyday tasks, such as search results and product recommendations [35], all the way to high-stakes decisions such as mortgage approval [17], job applicant filtering [25], and pre-trial risk assessment of criminals [55]. But such automated predictions are only as good as the data that drives them. Recent work has shown that inherent biases are common in data [5], and data-driven systems commonly demonstrate unfair and discriminatory behavior [8, 55, 69, 76].

It is natural that data management research has shown growing interest in the topic of fairness over applications related to ranking, data synthesis, result diversification, and others [2–4, 32, 50, 73, 81]. However, much of this work does not target prediction systems directly. In fact, a relatively small portion of the fairness literature within the data management community has directly targeted *classification* [26, 54, 69], one of the most important and heavily-studied supervised machine learning tasks that drives many broadly-used prediction systems. In contrast, machine learning research has rapidly produced a large body of work on the problem of improving fairness in classification.

In this paper, we closely study and evaluate existing work on fair classification, across different research communities, with two primary objectives: (1) to highlight data management aspects of existing work, such as efficiency and scalability, which are often overlooked in other communities, and (2) to produce a deeper understanding of tradeoffs and challenges across various approaches, creating guidelines for where data management solutions are more likely to have impact. We proceed to provide more detailed background on the problem of fair classification and existing approaches, we state the scope of our work and contrast with prior evaluation and analysis research, and, finally, we list our contributions.

Background on fair classification. Given the values of the predictive attributes of an entity, the task of a classifier is to predict which class, among a set of predefined classes, that entity belongs to. Classifiers typically focus on maximizing their *correctness*, which measures how well the predictions match the ground truth. To that end, they try to minimize the prediction error over a subset (validation set) of the available labeled data. Since both the training and the validation sets are drawn from the same source, the trained classifier naturally prioritizes the minimization of prediction error over the over-represented (majority) groups within the source data, and, thus, performs better for entities that belong to those groups. However, this may result in poor prediction performance over the under-represented (minority) groups. Moreover, as all data-driven approaches, classifiers also suffer from the general phenomenon of "garbage-in, garbage-out": if the data contains some inherent bias, the model will also reflect or even exacerbate it. Thus, traditional learning models may discriminate in two ways: (1) they make more incorrect predictions over the minority groups than the majority groups, and (2) they learn (and replicate) training data biases.

Consider COMPAS, a risk assessment system that can predict recidivism (the tendency to reoffense) in convicted criminals and is used by the U.S. courts to classify defendants as high- or low-risk according to their likelihood of recidivating within 2 years of initial assessment [28]. COMPAS achieves nearly 70% accuracy [23], which is a well-known metric to quantify a classifier's correctness. In 2014, a detailed analysis of COMPAS revealed some very troubling findings: (1) black defendants are twice more likely than white defendants to be predicted as high-risk, and (2) white reoffenders are predicted low-risk almost twice as often as black reoffenders [55]. While COMPAS' overall accuracy was similar over both groups (67% for black and 69% for white), its mistakes affected the two groups disproportionately. Furthermore, COMPAS was criticized for exacerbating societal bias due to its usage of historical arrest data in training, despite certain populations being proven to be more policed than others [67]. This is not an isolated incident; many other incidents of classifier discrimination have pointed towards racial [8], gender [69], and other forms of discrimination and unfairness [76].

The pervasiveness of examples of discriminatory behavior in prediction systems indicates that *fairness* should be an important objective in classification. In recent years, study of fair classification has garnered significant interest across multiple disciplines [16, 26, 37, 69, 85], and a multitude of approaches and notions of fairness have emerged [61, 78]. We consider two principal dimensions in characterizing the work in this domain: (1) the targeted notion of fairness, and (2) the stage—before, during, or after training—when fairness-enforcing mechanisms are applied.

Fairness notions and mechanisms. Specifying what is fair is non-trivial: the proper definition of fairness is often driven by application-specific and even legal considerations. There is a large number of fairness definitions [61, 78], and new ones continue to emerge. Some fairness notions capture if *individuals* are treated fairly, while others quantify fair treatment of a *group* (e.g., people of certain race or gender). Further, some notions measure discrimination through *causal* association among attributes of interest (e.g., race and prediction), while others study non-causal associations. The mechanism to quantify fairness also varies: some notions rely on *observational* data, while others apply *interventional* techniques. To add further complexity, recent studies show that some fairness notions are incompatible with others and cannot be enforced simultaneously [22].

Fairness-enforcing stage. Existing methods in fair classification operate in one of three possible stages. *Pre-processing* approaches attempt to repair biases in the data *before* the data is used to train a classifier [14, 26, 40, 69, 88]. Data management research in fair classification has typically focused on the pre-processing stage. In contrast, the machine learning community largely explored *in-processing* approaches, which alter the learning procedure used by the classifier [16, 44, 74, 83, 85, 87], and *post-processing* approaches, which alter the classifier predictions to ensure fairness [37, 42, 64].

Scope of our work. We present a thorough empirical evaluation of 13 fair classification approaches and some of their variants, resulting in 18 different approaches, across the aspects of *correctness*, *fairness*, *efficiency*, *scalability*, and *stability*. We selected approaches that target a representative variety of fairness definitions and span all three (pre, in, and post) fairness-enforcing stages. In general, there is no one-size-fits-all solution when it comes to choosing the best fair approach and the choice is application-specific. However, our evaluation has two main objectives: (1) to highlight issues of efficiency and scalability, which are often overlooked in other communities, and (2) to produce a deeper understanding of tradeoffs and challenges across various approaches, creating guidelines for where data management solutions are more likely to have impact. To the best of our knowledge, this is the first study and evaluation of fair classification approaches through a data management lens.

Other evaluation and analysis work on fair classification. Prior work on the evaluation of fair classifiers has had a more narrow scope than ours. Friedler et al. [30] compare variations of 4 fair approaches over 5 fairness metrics, while Jones et al. [39] evaluate variations of 6 fair approaches over 3 fairness metrics. Further, these evaluation studies do not examine runtime performance aspects, such as scalability, and do not include post-processing approaches or individual fairness metrics in their analysis.

AI Fairness 360 [6] is an extensible toolkit that tests 11 fair approaches on 7 fairness metrics, but it is not designed for comparative analysis of approaches and does not cover efficiency, scalability, and stability of classifiers. Other works [31, 75] provide general frameworks to evaluate approaches on some specific fairness metric, but are not extendable for evaluating over multiple metrics. Lastly, there are surveys that discuss fair approaches available in the literature [15, 59], but they do not evaluate them empirically.

Contributions. In this paper, we make the following contributions:

- We provide a new and informative categorization of 26 existing fairness notions, based on the high-level aspects of granularity, association, methodology, and requirements. We discuss their implications, tradeoffs, and limitations, and justify the choices of metrics for our evaluation. (Section 2)
- We provide an overview of 13 fair classification approaches and several variants. We select 5 *pre-processing* [14, 26, 40, 69, 88], 5 *in-processing* [16, 44, 74, 83, 85, 87], and 3 *post-processing* approaches [37, 42, 64] for our evaluation. (Section 3)
- We evaluate a total of 18 variants of fair classification techniques with respect to 4 correctness and 5 fairness metrics over 4 real-world datasets including Adult [49] and COMPAS [55]. Our evaluation provides interesting insights regarding the trends in fairness-correctness tradeoffs. (Section 4.2)
- Our runtime evaluation indicates that post-processing approaches are generally most efficient and scalable. However, their efficiency and scalability are due to the simplicity of their mechanism, which limits their capacity of balancing correctness-fairness tradeoffs. In contrast, pre- and in-processing approaches generally incur higher runtimes, but offer more flexibility in controlling correctness-fairness tradeoffs. With respect to scalability, pre-processing approaches (which have been the focus of the related data management literature) tend to be most affected by the number of attributes, while in-processing approaches have worse response to increasing dataset size. (Section 4.3)
- To evaluate stability, we measure the variance in correctness and fairness over different partitions of the training data. Our findings show that all evaluated approaches are generally stable and high-variance behavior is rare. (Section 4.4)
- Finally, based on the insights from our evaluation, we discuss general guidelines towards selecting suitable fair classification approaches in different settings, and highlight possible areas where data management solutions can be most impactful. (Section 5)

2 EVALUATION METRICS

In this section, we introduce the metrics that we use to measure the correctness and fairness of the evaluated techniques. We start with some basic notations related to the concepts of binary classification. Next, we proceed to describe the two types of evaluation metrics and the rationale behind our choices.

Basic notations.

Let $\mathcal{D}$ be an annotated dataset with the schema $(\mathbb{X}, S; Y)$, where $\mathbb{X}$ denotes a set of attributes that describe each tuple or individual in the dataset $\mathcal{D}$, S denotes a sensitive attribute, and Y denotes the annotation (ground-truth class label). Without loss of generality, we assume that S is binary, i.e., $\mathbf{Dom}(S) = \{0, 1\}$, where 1 indicates a *privileged* and 0 indicates an *unprivileged* group. We use S_t to denote the particular sensitive attribute assignment of a tuple $t \in \mathcal{D}$.

Notation	Description
$\mathbb{X}$	A set of attributes
X, $\mathbf{Dom}(X)$	A single attribute X and its value domain
S	A sensitive attribute
Y	Attribute denoting the ground-truth class label
$\mathcal{D}$	An annotated dataset with the schema $(\mathbb{X}, S; Y)$
$f(\mathbb{X}) \rightarrow \hat{Y}$	A binary classifier
$\hat{Y}$	Attribute that denotes the predicted class label
S_t	Value of the sensitive attribute S for tuple $t \in \mathcal{D}$
Y_t, $\hat{Y}_t$	Ground-truth and predicted class labels for tuple $t \in \mathcal{D}$

Figure 1: Summary of notations.

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	True Positive (*TP*) $TPR = Pr(\hat{Y}{=}1 \mid Y{=}1)$	False Positive (*FP*) $FPR = Pr(\hat{Y}{=}1 \mid Y{=}0)$
$\hat{Y} = 0$	False Negative (*FN*) $FNR = Pr(\hat{Y}{=}0 \mid Y{=}1)$	True Negative (*TN*) $TNR = Pr(\hat{Y}{=}0 \mid Y{=}0)$

Figure 2: Confusion matrix for predictions of a binary classifier.

Metric	Definition	Range	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	[0, 1]	Accuracy = 1 → completely correct Accuracy = 0 → completely incorrect
Precision	$\frac{TP}{TP+FP}$	[0, 1]	Precision = 1 → completely correct Precision = 0 → completely incorrect
Recall	$\frac{TP}{TP+FN}$	[0, 1]	Recall = 1 → completely correct Recall = 0 → completely incorrect
F_1-score	$\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$	[0, 1]	F_1-score = 1 → completely correct F_1-score = 0 → completely incorrect

Figure 3: List of correctness metrics used in our evaluation.

We denote a binary classification task $f : f(\mathbb{X}) \rightarrow \hat{Y}$, where $\hat{Y}$ denotes the *predicted* class label ($\mathbf{Dom}(Y) = \mathbf{Dom}(\hat{Y}) = \{0, 1\}$). Without loss of generality, we interpret 1 as a favorable (positive) prediction and 0 as an unfavorable (negative) prediction. Y_t and $\hat{Y}_t$ denote the ground-truth and predicted class label for t, respectively. We summarize the notations used in the paper in Figure 1.

2.1 Correctness

The correctness of a binary classifier measures how well its predictions match the ground truth. Given a dataset $\mathcal{D}$ and a binary classifier f, we profile f's predictions on $\mathcal{D}$ using the statistics depicted in Figure 2, where *TP*, *TN*, *FP*, and *FN* are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Among the positive tuples ($Y = 1$), the true positive rate (*TPR*) is the fraction of tuples that are *correctly* predicted as positive and the false negative rate (*FNR*) is the fraction of tuples that are *incorrectly* predicted as negative.

Similarly, among the negative tuples ($Y = 0$), the true negative rate (*TNR*) is the fraction of tuples that are *correctly* predicted as negative and the false positive rate (*FPR*) is the fraction of tuples that are *incorrectly* predicted as positive.

Metrics. In our evaluation, we measure correctness using the metrics in Figure 3, which are widely-accepted and well-studied in the literature [52]. Intuitively, *accuracy* captures the overall correctness of the predictions made by a classifier; *precision* captures "preciseness", i.e., the fraction of positive predictions that are correctly predicted as positive; and *recall* captures "coverage", i.e., the fraction of positive tuples that are correctly predicted as positive. The F_1*-score* is the harmonic mean of precision and recall. While accuracy is an effective correctness metric for classifiers operating on datasets with a balanced class distribution, it can be misleading when the dataset is imbalanced, a frequently-observed scenario in real-world datasets. In such cases, precision, recall, and F_1-score, together, provide better understanding of correctness.

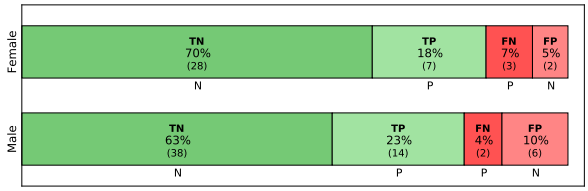


Figure 4: Prediction statistics over 100 applicants, grouped by gender: 60 male (bottom) and 40 female (top). The ground truth (positives as *P* and negatives as *N*) is indicated below each segment.

2.2 Fairness

Fairness in classifier predictions typically targets sensitive attributes, such as gender, race, etc. We highlight ways in which classifier predictions can discriminate through an example.

Example 1. *Consider a model of university admissions that aims to offer admission to highly-qualified students. The admissions committee automates the admission process by training a binary classifier over historical admissions data. Female students are historically underrepresented at this university, making up 40% of the student body; so, we designate males as the privileged group ($S = 1$), and females as the unprivileged group ($S = 0$). After training, the classifier achieves 87% accuracy and 78% F_1-score over the training data. Figure 4 summarizes the prediction-related statistics for both groups. Although the classifier is satisfactory in terms of correctness, it is not fair across gender. Specifically, we observe two ways females are being discriminated:*

- *(Discrimination-1) The fraction of females predicted as highly-qualified (positive) is $\frac{7+2}{40} \approx 23\%$, which is significantly lower than the fraction of males predicted as highly-qualified ($\frac{14+6}{60} \approx 33\%$).*
- *(Discrimination-2) The true positive rate for females is $\frac{TP}{TP+FN} = \frac{7}{7+3} = 70\%$, which is significantly lower than that of the males ($\frac{TP}{TP+FN} = \frac{14}{14+2} \approx 88\%$).*

Example 1 outlines two ways a classifier can be unfair despite having reasonable accuracy in predictions overall. Discrimination-1 highlights how a group can receive an unfair advantage (or disadvantage) if the proportion of positive and negative predictions differs across groups. On the other hand, Discrimination-2 indicates how predictions can disadvantage a group if the correctness of predictions (e.g., *TPR*) differs across groups. These disparities are very common in real-world scenarios [21] and underscore the need for ensuring fairness in classification.

2.2.1 Fairness Notions. Fairness is not entirely objective, and societal requirements and legal principles often demand different characterizations. Fairness is also a relatively new concern within the research community. Consequently, a large number of different fairness definitions have emerged, along with a variety of metrics to quantify them. Figure 5 presents a list of 26 fairness notions and their corresponding metrics that have been studied in the literature.

Fairness notion	Metric	Granularity		Association		Methodology		Additional requirements				
		group	individual	causal	non-causal	observational	interventional	ground truth	prediction probability	causality model	resolving attribute	similarity metric
demographic parity† [24]	disparate impact [85], CV score [12]	✓			✓	✓						
conditional statistical parity [22]	conditional statistical parity	✓			✓	✓						
intersectional fairness [29]	differential fairness	✓			✓	✓						
conditional accuracy equality [8]	false discovery/omission rate parity	✓			✓	✓		✓				
predictive parity [20]	false discovery rate parity	✓			✓	✓		✓				
overall accuracy equality [8]	balanced classification rate [30]	✓			✓	✓		✓				
treatment equality [8]	ratio of false negative and false positive	✓			✓	✓		✓				
equalized odds [37]	true positive/negative rate balance	✓			✓	✓		✓				
equal opportunity‡ [37]	true negative rate balance	✓			✓	✓		✓				
resilience to random bias [27]	resilience to random bias	✓			✓	✓		✓				
preference-based fairness [84]	group benefit	✓			✓	✓		✓				
calibration [20]	calibration	✓			✓	✓		✓	✓			
calibration within groups [48]	well calibration	✓			✓	✓		✓	✓			
positive class balance [48]	fairness to positive class	✓			✓	✓		✓	✓			
negative class balance [48]	fairness to negative class	✓			✓	✓		✓	✓			
causal discrimination [31]	causal discrimination		✓	✓			✓					
counterfactual fairness [51]	counterfactual effect [80]		✓	✓			✓			✓		
path-specific fairness [60]	natural direct effects	✓		✓			✓			✓		
path-specific counterfactuals [80]	path-specific effect, counterfactual effect		✓	✓			✓			✓		
fair causal inference [58]	estimation of heterogeneous effects [38]		✓	✓			✓			✓		
proxy fairness [45]	proxy fairness	✓		✓			✓			✓		
unresolved discrimination [45]	causal risk difference [66]	✓		✓		✓					✓	
interventional/justifiable fairness [69]	ratio of observable discrimination	✓		✓		✓					✓	
metric multifairness [47]	metric multifairness		✓		✓	✓						✓
fairness through awareness [24]	fairness through awareness		✓		✓	✓						✓
fairness through unawareness [51]	Kusner et al. [51]		✓		✓							

Figure 5: List of fairness definitions and their corresponding metrics in the literature. We list key properties of notions such as the level of granularity (group or individual), type of association considered between attributes (causal or non-causal), and technique of measuring fairness (observational or interventional). All notions require knowledge of the sensitive attributes and the predictions made by the classifier. Some fairness definitions rely on additional requirements that are shown in the rightmost four columns. For our evaluation, we choose five fairness metrics (Figure 6) that cover the highlighted definitions. (†also known as statistical parity; ‡also known as predictive equality)

We offer a new categorization of these notions based on their granularity, association, methodology, and requirements they impose:

Granularity. We classify fairness notions into two categories based on the granularity of their target: *group* fairness characterizes if any demographic group, collectively, is being discriminated against; *individual* fairness determines if similar individuals are treated similarly, regardless of the values of the sensitive attribute.

Association. All notions characterize fairness by investigating the existence of some association between the sensitive attribute and the prediction. The type of association can be either *causal*, which analyzes the source of discrimination through the causal relationships among the attributes, or *non-causal*, which includes observed statistical correlations among the attributes.

Methodology. We identify an important methodological distinction in fairness notions: most definitions are based on measurements over *observational* data, while others apply *interventional* methods to generate what-if scenarios and measure fairness based on predictions of those scenarios.

Additional requirements. All notions require information on the sensitive attribute and the classifier predictions. Some notions impose additional requirements, such as causality models that capture the causal relationships, resolving attributes that depend on the sensitive attribute in non-discriminatory ways, similarity metric between individuals, etc.

2.2.2 Fairness Metrics. While Figure 5 highlights the wide range of proposed fairness notions, Friedler et al. [30] have shown that a large number of metrics (and their notions) strongly correlate with one another, and, thus, are highly redundant. For our evaluation, we carefully selected five fairness metrics (Figure 6) that are most prevalent in the literature and that capture commonly observed discriminations in binary classification [20]. Moreover, we ensured that our selected metrics cover all categories in our classification, including group- and individual-level fairness, causal and non-causal associations, and observational and interventional methods (highlighted rows in Figure 5). We proceed to describe our chosen metrics.

Disparate Impact (DI) is a group, non-causal, and observational metric. It quantifies demographic parity [24], a fairness notion that states that positive predictions should be independent of the sensitive attribute. To measure demographic parity, *DI* computes the ratio of empirical probabilities of receiving positive predictions between the unprivileged and the privileged groups.

$$DI = \frac{Pr(\hat{Y} = 1 \mid S = 0)}{Pr(\hat{Y} = 1 \mid S = 1)}$$

DI lies in the range $[0, \infty)$. $DI = 1$ denotes perfect demographic parity. $DI < 1$ indicates that the classifier favors the privileged group and $DI > 1$ means the opposite. In Example 1, $DI = \frac{9/40}{20/60} = 0.67$, which suggests that positive predictions are not independent

Metric	Definition	Fairness notion	Range	Interpretation
Disparate Impact (*DI*) [26]	$\frac{Pr(\hat{Y}=1 \mid S=0)}{Pr(\hat{Y}=1 \mid S=1)}$	demographic parity	$[0, \infty)$	$DI = 1 \rightarrow$ completely fair $DI = 0 \rightarrow$ completely unfair $DI = \infty \rightarrow$ completely unfair
True Positive Rate Balance (*TPRB*) [37]	$Pr(\hat{Y}=1 \mid Y=1, S=1) - Pr(\hat{Y}=1 \mid Y=1, S=0)$	equalized odds	[-1, 1]	$\lvert TPRB \rvert = 0 \rightarrow$ completely fair $\lvert TPRB \rvert = 1 \rightarrow$ completely unfair
True Negative Rate Balance (*TNRB*) [37]	$Pr(\hat{Y}=0 \mid Y=0, S=1) - Pr(\hat{Y}=0 \mid Y=0, S=0)$	equalized odds	[-1, 1]	$\lvert TNRB \rvert = 0 \rightarrow$ completely fair $\lvert TNRB \rvert = 1 \rightarrow$ completely unfair
Causal Discrimination (*CD*) [31]	$\frac{\lvert Q \rvert}{\lvert \mathcal{D} \rvert}$, given $Q = \{a \in \mathcal{D} \mid \exists b : \mathbb{X}_a{=}\mathbb{X}_b \wedge S_a{\neq}S_b \wedge \hat{Y}_a{\neq}\hat{Y}_b\}$	causal discrimination	[0, 1]	$CD = 0 \rightarrow$ completely fair $CD = 1 \rightarrow$ completely unfair
Causal Risk Difference (*CRD*) [66]	$\frac{\sum_{t \in \mathcal{D}} w(t) \cdot [\![S_t{=}1 \wedge \hat{Y}_t{=}1]\!]}{\sum_{t \in \mathcal{D}} w(t) \cdot [\![S_t{=}1]\!]} - Pr(\hat{Y}{=}1 \mid S{=}0)$, given $w(t){=}\frac{Pr(S_t{=}0 \mid \mathbb{R})}{1-Pr(S_t{=}0 \mid \mathbb{R})}$, $\mathbb{R} \subseteq \mathbb{X}$ is a set of resolving attributes	unresolved discrimination	[-1, 1]	$\lvert CRD \rvert = 0 \rightarrow$ completely fair $\lvert CRD \rvert = 1 \rightarrow$ completely unfair

Figure 6: List of fairness metrics we use to evaluate fair classification approaches. These metrics capture group- and individual-level discrimination; and effectively contrast between causal and non-causal associations, observational and interventional techniques.

of gender as males have higher probability to receive positive predictions than females. This is indicative of Discrimination-1: the fraction of females being granted admission is much lower than males.

True Positive Rate Balance (TPRB) and True Negative Rate Balance (TNRB) are two group, non-causal, and observational metrics. They measure discrimination as the difference in *TPR* and *TNR*, respectively, between the privileged and unprivileged groups.

$$TPRB = Pr(\hat{Y}{=}1 \mid Y{=}1, S{=}1) - Pr(\hat{Y}{=}1 \mid Y{=}1, S{=}0)$$
$$TNRB = Pr(\hat{Y}{=}0 \mid Y{=}0, S{=}1) - Pr(\hat{Y}{=}0 \mid Y{=}0, S{=}0)$$

Both *TPRB* and *TNRB* lie in the range $[-1, 1]$. These two metrics, together, measure equalized odds [37], which states that prediction statistics (e.g., *TPR* and *TNR*) should be similar across the privileged and the unprivileged groups. Perfect equalized odds is achieved when *TPRB* and *TNRB* are 0, as the classifier performs equally well for both groups. A positive value in either of the two metrics indicates that the classifier tends to misclassify the unprivileged group more. In Example 1, $TPRB = \frac{14}{16} - \frac{7}{10} = 0.18$ and $TNRB = \frac{38}{44} - \frac{28}{30} = -0.07$. The high positive value of *TPRB* indicates Discrimination-2: the *TPR* of females is much lower than males.

Causal Discrimination (CD) [31] is an individual, causal, and interventional metric. It allows us to determine both the classifier's discrimination with respect to individuals and the causal influence of the sensitive attribute. Specifically, *CD* is the fraction of tuples for which, changing the sensitive attribute causes a change in the prediction, compared to otherwise identical data points. Suppose that Q is the set of such tuples, defined as $Q = \{a \in \mathcal{D} \mid \exists b : \mathbb{X}_a{=}\mathbb{X}_b \wedge S_a{\neq}S_b \wedge \hat{Y}_a{\neq}\hat{Y}_b\}$; then $CD = \frac{|Q|}{|\mathcal{D}|}$. *CD* lies in the range $[0, 1]$ and $CD = 0$ indicates that there exists no data point for which the sensitive attribute is the cause of discrimination.

Example 2. *Consider 7 university applicants shown in Figure 7. To measure CD, we intervene on the sensitive attribute (*`gender`*) of each tuple while keeping rest of the attributes intact, and re-evaluate the classifier on the altered tuples. Suppose that the prediction for t_6 changes from 0 to 1 when t_6's* `gender` *is altered from* `Female` *to* `Male`*, and that predictions do not change for any other tuples. Then, $CD = \frac{1}{7} = 0.14$, indicating that 14% of the applicants are directly discriminated because of their gender.*

	$\mathbb{X}$			S	$\hat{Y}$
id	SAT	dept_choice	rank	gender	admitted
t_1	1200	Physics	11	Male	0
t_2	1350	Mathematics	03	Male	1
t_3	1105	Physics	09	Female	1
t_4	1410	Mathematics	03	Female	1
t_5	1130	Marketing	10	Male	1
t_6	1290	Mathematics	12	Female	0
t_7	1210	Marketing	11	Male	1

Figure 7: Sample data for 7 university applicants.

The formal definition of *CD* requires interventions on all possible data points in the domain of attributes, but practical heuristics limit interventions to smaller datasets of interest [31].

Causal Risk Difference (CRD) [66] is a group, causal, and observational metric. It quantifies discrimination by measuring the difference in probability of positive prediction between the privileged and the unprivileged groups, accounting for the confounding effects of the resolving attributes. *CRD* is computed from observational data using the following steps:

- To filter out the confounding effects, *CRD* first computes a *propensity score*, as the conditional probability of belonging to the unprivileged group given a set of resolving attributes $\mathbb{R}$, for each tuple: $propScore(t) = Pr(S = 0 \mid \mathbb{R})$.
- The propensity score is then used to assign each tuple a weight: $w(t) = \frac{propScore(t)}{1-propScore(t)}$. Tuples with propensity scores > 0.5 will have weights > 1.
- Finally, *CRD* is formally expressed as:

$$CRD = \frac{\sum_{t \in \mathcal{D}} w(t) \cdot [\![S_t = 1 \wedge \hat{Y}_t = 1]\!]}{\sum_{t \in \mathcal{D}} w(t) \cdot [\![S_t = 1]\!]} - Pr(\hat{Y}{=}1 \mid S{=}0)$$

CRD lies in the range $[-1, 1]$ and $CRD = 0$ implies no discrimination, if we account for the effects of the resolving attributes.

Example 3. *Consider the applicants in Figure 7. Further, suppose that females tend to apply to the Physics and Mathematics departments, and that these two departments have low acceptance rates. By setting* $\mathbf{R}$ *=* {`dept_choice`}*, we get higher propensity scores for tuples in Mathematics and Physics, i.e., those applicants are more likely to be female, and these tuples contribute more to CRD. Based on the data of Figure 7, we get that $w(t_1) = w(t_3) = 1$, $w(t_2) = w(t_4) = w(t_6) = 2$,*

and $w(t_5) = w(t_7) = 0$. *Computing the causal risk difference, we get* $CRD = \frac{2+0+0}{1+2+0+0} - \frac{2}{3} = 0$. *In this case, CRD indicates that there is no discrimination when the choice of department is accounted for.*

Discussion on metric choices. DI, *TPRB*, and *TNRB* address group-level and non-causal discrimination. This means they do not capture discrimination against individuals that may be masked in group aggregates, and they do not account for confounding factors in the data. On the other hand, *CD* captures individual discrimination, and *CRD* can remove confounding effects by determining if the apparent discrimination found in the observational data is explainable through *resolving* attributes, i.e., attributes that are dependent on or are implicitly influenced by the sensitive attribute in non-discriminatory ways (e.g., choice of department in Example 3).

Other causal notions (Figure 5) can also address the limitations of non-causal metrics. However, they typically rely on graphical or mathematical causality models to express the cause-and-effect relationships among attributes. We exclude them from our evaluation, because determining such causality models requires making strong assumptions about the problem setting, which is often impractical [63]. Further, we do not include non-causal individual level metrics, because they require a similarity measure between individuals, which requires domain expertise.

3 FAIR CLASSIFICATION APPROACHES

Fair classification techniques vary in the fairness notions they target and the mechanisms they employ. We categorize approaches based on the stage when fairness-enforcing mechanisms are applied. (1) *Pre-processing* approaches attempt to repair biases in the data before training; (2) *in-processing* approaches modify the learning procedure to include fairness considerations; finally, (3) *post-processing* approaches modify the predictions made by the classifier. Figure 8 contains an overview of the fair approaches that we choose for our evaluation. We proceed to provide a high-level description of the approaches in each category, highlighting their similarities and differences. (More details are in the Appendix).

Pre-processing approaches are motivated from the fact that machine learning techniques are data-driven and the predictions of a classifier reflect the trend and biases of the training data. Data management research most naturally fits in this category. These approaches modify the data before training to remove biases, which subsequently ensures that the predictions of a learned classifier satisfy the target notion of fairness. The main advantage of pre-processing is that it is model-agnostic, which allows flexibility in choosing the classifiers based on the application requirements. However, since pre-processing happens *before* training and does not have access to the predictions, these approaches are limited in the number of notions they can support and does not always come with provable guarantees of fairness.

Demographic parity is one of the most widely used fairness notions among pre-processing approaches that enforce non-causal notions [12, 14, 26, 40, 41]. We evaluate three pre-processing approaches that enforce demographic parity. Kam-Cal [40] resamples the training data $\mathcal{D}$ with a weighted sampling technique to remove dependencies between the sensitive attribute S and the target attribute Y. In contrast, Calmon [14] and Feld [26] directly modify the data. Calmon modifies both $\mathbb{X}$ and Y to reduce dependency between Y and S, while preventing major distortion of the joint data distribution and significant change of the attribute values. Feld argues that a model that learns only from the attributes that are independent of S is likely to make predictions that are independent of S as well. To this end, Feld modifies $\mathbb{X}$ in a way that ensures that the marginal distribution of each individual attribute is indistinguishable across the sensitive groups. Feld controls the extent of repair with a parameter $\lambda \in [0, 1]$. In our evaluation, we choose two values of λ (1.0 and 0.6) to highlight its impact on the performance.

We also evaluate two pre-processing approaches that do not target demographic parity. Zha-Wu [88] enforces path-specific fairness by modifying Y such that all causal influence of S over Y is removed. To this end, it learns a graphical causal model over $\mathcal{D}$ to discover (direct and indirect) causal associations between Y and S. Salimi [69] enforces justifiable fairness, which prohibits causal dependence between S and $\hat{Y}$, except through admissible attributes. Salimi does not depend on the causal model; it translates justifiable fairness to an integrity constraint over $\mathcal{D}$, and minimally repairs $\mathcal{D}$ using tuple insertion and deletion. The repair problem is reduced to two NP-hard problems: weighted maximum satisfiability (MaxSAT) [9] and matrix factorization (MatFac) [56].

In-processing approaches are most favored by the machine learning community [16, 44, 85, 87] and the majority of the fair classification approaches fall under this category. In-processing takes place within the training stage and fairness is typically added as a constraint to the classifier's objective function (that maximizes correctness). The advantage of in-processing lies precisely in the ability to adjust the classification objective to address fairness requirements directly, and, thus has the potential to provide guarantees. However, in-processing techniques are model-specific and require re-implementation of the learning algorithms to include the fairness constraints. This hinges on the assumption that the model is replaceable or modifiable, which may not always be the case.

We evaluate five in-processing approaches and their variants. Zafar [83, 85] proposes two approaches to enforce demographic parity and equalized odds, which utilize tuples' distance from the decision boundary as a proxy of $\hat{Y}$ to model fairness violations, and translate the fairness notion to convex functions of the classifier parameters. Zafar then solves the resulting constrained optimization problem that either maximizes prediction accuracy under fairness constraints, or minimizes fairness violation under constraints on accuracy compromise. Zha-Le [87] enforces equalized odds through adversarial learning: a fair classifier is trained, such that an adversary cannot predict S from the knowledge of Y and $\hat{Y}$. Kearns [44] interpolates between group and individual fairness: it guarantees fairness for a large set of subgroups within the training population by constructing constraints that restrict the amount of fairness violation in each group. Celis [16] and Thomas [74] provide general frameworks that accommodate a large number of notions. Celis reduces all fairness notions to linear forms and solves the corresponding constrained optimization problem to minimize prediction error under fairness constraints. Given a fairness notion, Thomas utilizes concentration inequalities to compute the worst possible fairness violation a classifier can incur, and then selects classifier parameters for which this violation is within an allowable threshold.

Stage	Approach	Fairness notion(s)	Key mechanism	Evaluated version(s)
pre	Kam-Cal [40]	demographic parity	Apply weighted resampling over tuples in $\mathcal{D}$ to remove dependency between S and Y.	• Kam-Cal^{DP}
	Feld [26]	demographic parity	Repair each $X \in \mathbb{X}$ independently s.t. X's marginal distribution is indistinguishable across sensitive groups. A user-defined parameter $\lambda \in [0, 1]$ specifies degree of repair.	• Feld$^{DP}_{\lambda=1.0}$ (Full repair with $\lambda = 1.0$) • Feld$^{DP}_{\lambda=0.6}$ (Partial repair with $\lambda = 0.6$)
	Calmon [14]	demographic parity	Modify $\mathbb{X}$ and Y to reduce dependency between Y and S, while preventing major distortion of the joint data distribution and significant change of the attribute values.	• Calmon^{DP}
	Zha-Wu [88]	path-specific fairness	Exploit a (learned) causal model over the attributes to discover (direct and indirect) causal association between Y and S. Modify Y to remove such causal association.	• Zha-Wu^{PSF}
	Salimi [69]	justifiable fairness	Mark attributes as *admissible* (A)—allowed to have causal association—or *inadmissible* (I)—prohibited to have causal association—with Y; repair $\mathcal{D}$ to ensure that Y is conditionally independent of I, given A. Reduce the repair problem to known problems.	• Salimi$^{JF}_{MaxSAT}$ (Weighted maximum satisfiability) • Salimi$^{JF}_{MatFac}$ (Matrix factorization)
in	Zafar [83, 85]	demographic parity equalized odds	Use tuple t's distance from the decision boundary as a proxy of $\hat{Y}_t$. Model fairness violation by the correlation between this distance and S over all tuples in $\mathcal{D}$. Solve variations of constrained optimization problem that either maximizes prediction accuracy under constraint on maximum fairness violation, or minimizes fairness violation under constraint on maximum allowable accuracy compromise.	• Zafar$^{DP}_{Fair}$ (Maximize accuracy under constraint on demographic parity) • Zafar$^{DP}_{Acc}$ (Maximize demographic parity under constraint on accuracy) • Zafar$^{EO}_{Fair}$ (Same as Zafar$^{DP}_{Fair}$, but use misclassified tuples only)
	Zha-Le [87]	equalized odds	Learn classifier $f : f(\mathbb{X}, S) \rightarrow \hat{Y}$ and adversary $\alpha : \alpha(Y, \hat{Y}) \rightarrow \hat{S}$ together. Enforce fairness by ensuring that α cannot infer S from Y and $\hat{Y}$.	• Zha-Le^{EO}
	Kearns [44]	demographic parity predictive equality	Use sensitive attribute(s) to construct a set of subgroups. Define fairness constraint s.t. the probability of positive outcomes (demographic parity) or *FPR* (predictive equality) of each subgroup matches that of the overall population.	• Kearns^{PE} (For subgroups $\{\mathcal{D}_1, \mathcal{D}_2, \dots\}$ where each $\mathcal{D}_i \subset \mathcal{D}$, ensure that $\forall \mathcal{D}_i$, $FPR(\mathcal{D}_i) \approx FPR(\mathcal{D})$)
	Celis [16]	equalized odds demographic parity predictive parity cond. acc. equality	Unify multiple fairness notions in a general framework by converting the fairness constraints to a linear form. Solve the corresponding linear constrained optimization problem s.t. prediction error is minimized under fairness constraints.	• Celis^{PP} (Enforce $Pr(Y{=}0 \mid \hat{Y}{=}1, S{=}0) \approx Pr(Y{=}0 \mid \hat{Y}{=}1, S{=}1)$)
	Thomas [74]	demographic parity equalized odds equal opportunity predictive equality	Compute worst possible fairness violation a classifier can incur for a set of parameters and pick parameters for which this worst possible violation is within an allowable threshold.	• Thomas^{DP} (Enforce demographic parity) • Thomas^{EO} (Enforce equalized odds)
post	Kam-Kar [42]	demographic parity	Modify $\hat{Y}$ for tuples close to the decision boundary (i.e., subject to low prediction confidence) s.t. the probability of positive outcome is similar across sensitive groups.	• Kam-Kar^{DP}
	Hardt [37]	equalized odds	Derive new predictor based on $\hat{Y}$ and S s.t. *TPR* and *TNR* are similar across sensitive groups.	• Hardt^{EO}
	Pleiss [64]	equal opportunity predictive equality	Modify $\hat{Y}$ for random tuples to equalize *TPR* (or *FPR*) across sensitive groups.	• Pleiss^{EOP} (Equalize *TPR*)

Figure 8: List of fair approaches, fairness notions they support, and high-level descriptions of the mechanisms they apply to ensure fairness. According to the stage of the classifier pipeline where fairness-enhancing mechanism is applied, these approaches are divided into three groups: (1) pre-processing, (2) in-processing, and (3) post-processing. In the rightmost column, we list the variations of each approach that we consider in our evaluation. We denote in the superscript the fairness notion that a specific variation is designed to support.

Post-processing approaches enforce fairness by manipulating the predictions made by an already-trained classifier. Like pre-processing, these approaches are also model-agnostic. Their benefit is that they do not require classifier retraining. However, since post-processing is applied in a late stage of the learning process, it offers less flexibility than pre- and in-processing.

We evaluate three post-processing approaches. Kam-Kar [42] modifies $\hat{Y}$ for tuples that are close to the decision boundary (i.e., the classifier has low prediction confidence for them), such that demographic parity is achieved across the sensitive groups. Hardt [37] enforces equalized odds by learning a new predictor derived from $\hat{Y}$ and S that equalizes *TPR* and *FPR* across the sensitive groups. Pleiss [64] enforces equal opportunity (equal *TPR* across the sensitive groups) or predictive equality (equal *FPR* across the sensitive groups) while maintaining the consistency between the classifier's prediction probability for a class with the expected frequency of that class. To achieve this, it modifies $\hat{Y}$ for a random subset of tuples within the group with higher *TPR* (or lower *FPR*).

Other approaches. Beyond the ones we evaluate, other fair classification approaches exist in the literature. Some are incorporated in the approaches we evaluate [12, 13, 41]. Others are empirically inferior [43], offer weaker guarantees [1, 65], do not offer a practical solution [62, 79], or do not apply to the classification setting [34, 53, 57, 70]. Some require additional information such as intermediate attributes [86], causal model [19, 46, 51, 60, 68], context-specific similarity metric between individuals and human judgments [24, 54], which are dataset-specific and hinge on domain knowledge.

4 EVALUATION AND ANALYSIS

In this section, we present results of our comparative evaluation over 18 variations of fair classification approaches as listed in Figure 8. The objectives of our performance evaluation are: (1) to contrast the effectiveness of fair classification approaches in enforcing fairness and observe correctness-fairness tradeoffs, i.e., the compromise in correctness to achieve fairness (Section 4.2), (2) to contrast efficiency and scalability of the fair classification approaches with varying dataset size and dimensionality (Section 4.3), and (3) to contrast stability (lack of variability) of these approaches over different partitions of the training data (Section 4.4). Our results affirm and extend previous results reported by the evaluated approaches.

Additionally, we present a comparative analysis, focusing on the stage dimension (pre, in, and post). Our analysis highlights findings that explain the behavior of fair approaches in different settings. For

example, we find that the impact of enforcing a specific fairness notion can be explained through the score of a fairness-unaware classifier for that notion: larger discrimination by the fairness-unaware classifier indicates that a fair approach that targets that notion will likely incur higher drop in accuracy. Further, we provide novel insights that underscore the strengths and weaknesses across pre-, in-, and post-processing approaches. We find that post-processing approaches are very efficient and scalable, but perform less well in the correctness-fairness dimensions; in contrast, pre-processing and in-processing approaches are generally less scalable with increasing data dimensionality and increasing data size, respectively, but handle the correctness-fairness tradeoff more flexibly.

We begin by providing details about our experimental settings: approaches we evaluate, their implementation details, metrics we use to evaluate the approaches, and the datasets we use. Then we proceed to present our empirical findings.

4.1 Experimental Settings

Approaches. We evaluated 18 variants based on 13 fair classification approaches (Figure 8). We limited our evaluation to variants with available implementations, as each variant typically requires non-trivial extension to the available codebases. Pre-processing approaches require the repaired data to be paired with a classifier to complete the model pipeline and we used logistic regression as the classifier. This is in line with the evaluations of the original papers as the use of logistic regression is common across all pre-processing approaches. Finally, to contrast the fairness-aware approaches against a fairness-unaware approach, we trained an unconstrained logistic regression classifier (LR) over each dataset.

System and implementation. We conducted the experiments on a machine equipped with Intel(R) Core(TM) i5-7200U CPU (2.71 GHz, Quad-Core) and 8 GB RAM, running on Windows 10 (version 1903) operating system. We collected some of the source codes from the authors' public repositories, some by contacting the authors, and the rest from the open source library AI Fairness 360 [6] (additional details are in the Appendix). All the approaches are implemented in Python. We implemented the fairness-unaware classifier LR using Scikit-learn (version 0.22.1) in Python 3.6. Implementations of all these approaches use a single-threaded environment, i.e., only one of the available processor cores is used. We implemented the evaluation script in Python 3.6.[1]

Metrics. We evaluated all approaches using four correctness metrics (Figure 3) and five fairness metrics (Figure 6). We normalize fairness metrics to share the same range, scale, and interpretation. We report $DI^* = \min(DI, \frac{1}{DI})$, which ensures that low fairness with respect to DI ($DI \to 0$ and $DI \to \infty$) is mapped to low values for DI^*. Further, we report $1 - |TPRB|$, $1 - |TNRB|$, $1 - CD$, and $1 - |CRD|$; this way, high discrimination with respect to, say, $TPRB$, maps to low fairness value in $1-|TPRB|$. Moreover, CD requires two parameters: a confidence fraction and an error-bound. We choose a confidence of 99% and error-bound of 1%, which implies that discrimination computed using CD is within 1% error margin of the actual discrimination with 99% confidence.

[1] https://github.com/maliha93/FairnessAnalysis

Dataset	Size (MB)	$\lvert\mathcal{D}\rvert$	$\lvert\mathbb{X}\rvert$	S	Sensitive groups		Target task
					Unprivileged	Privileged	
Adult	5.80	45,222	14	Sex	Female	Male	Income ≥ $50K
COMPAS	0.30	7,214	11	Race	African-American	Others	Risk of recidivism
German	0.06	1,000	9	Sex	Female	Male	Credit risk
Credit	2.50	20,651	26	Sex	Female	Male	Default on loan

Figure 9: Summary of the datasets. We choose our datasets to be varied in size, number of data points, number of attributes, and different instances of sensitive-attribute-based discrimination. We provide the target prediction tasks in the rightmost column.

Datasets. Our evaluation includes 4 real-world datasets, summarized in Figure 9. Each dataset contains varied degrees of real-world biases, allowing for the evaluation of the fair classification approaches against different scenarios. Furthermore, these datasets are well-studied in the fairness literature and are frequently used as benchmarks to evaluate fair classification approaches [30, 39, 59].

Adult [49] contains information about individuals from the 1994 US census. It contains records of more than 45,000 individuals and their information over 14 demographic and occupational attributes such as race, sex, education level, marital status, occupation, etc. The target task is to predict the income levels of individuals. Favorable/positive label ($Y = 1$) denotes high-income (income ≥ $50,000) and unfavorable/negative label ($Y = 0$) indicates low-income (income < $50,000). The percentage of high-income individuals in Adult is 24%. The dataset reflects historical gender-based income inequality: 11% of the females report high income, compared to 32% of the males. Hence, we choose `sex` as the sensitive attribute with `female` as the unprivileged and `male` as the privileged group.

COMPAS [55], compiled by ProPublica, contains criminal assessment information about defendants arrested in 2013-2014 and their assessment scores by the COMPAS recidivism tool [23]. It contains more than 7,200 data points and 11 attributes such as age, sex, prior arrest counts, charges pressed, etc. The target task is to predict whether an individual re-offends within two years of initial assessment. Positive label indicates that an individual does not recidivate and negative label indicates that an individual recidivates. 44% of the individuals in this dataset recidivate and the data contains racial bias: the percentage of re-offenders is much higher in African-Americans (51%), compared to others (39%). Hence, we select `race` as the sensitive attribute with `African-American` as the unprivileged and all other races as the privileged group.

German [33] contains 1,000 instances representing individuals applying for credit or loan to a bank, with attributes age, sex, type of job, credit information, etc. The target task is to predict credit risk. Positive label indicates low credit risk and negative label indicates high credit risk. Over the entire population, 70% are of low credit risk. This percentage is slightly lower for females than males: 65% vs 71%. Hence, we choose `sex` as the sensitive attribute with `female` as the unprivileged and `male` as the privileged group.

Credit [82] originated from a research aimed at predicting loan defaulting behavior of individuals in Taiwan. It contains information about more than 20,000 individuals over 24 attributes such as education, marital status, history of past payments, etc. The target

task is to predict whether an individual defaults on the next payment. Positive label represents timely payment and negative label indicates default. In this dataset, 67% do not default. The dataset is biased against females: 56% of females, compared to 75% of males, do not default. Hence, we choose `sex` as the sensitive attribute with `female` as the unprivileged and `male` as the privileged group.

Train-validation-test setting. The train-test split for each dataset was 70%-30% (using random selection) and we validated each classifier using 3-fold cross validation.

4.2 Correctness and Fairness

Figure 10 presents our correctness and fairness results over all approaches and metrics across the 4 datasets. Below, we discuss the key findings of this evaluation.

The fairness performance of fairness-unaware approaches influences the relative accuracy of fair approaches. Classifiers typically target accuracy as their optimization objective. Fair approaches, directly or indirectly, modify this objective to target both fairness and accuracy. When a fairness-unaware technique displays significantly different performance across different fairness metrics (e.g., low fairness wrt *DI* and high fairness wrt *TPRB*), this appears to translate to a significant difference in the accuracy of fair approaches that target these fairness metrics (higher accuracy drop for approaches that target *DI*, and lower drop for those that target *TPRB*).

Figure 10(a) demonstrates this scenario for Adult. LR trained on this dataset achieves high fairness in terms of *TPRB* and *TNRB*, but exhibits very low fairness in terms of *DI*. We observe that the approaches that optimize *DI* (such as Kam-Cal^{DP} and Calmon^{DP}) demonstrate a much larger accuracy drop than the approaches that target the equalized odds metrics (such as Zafar$^{EO}_{Fair}$, Zha-Le^{EO}, and Kearns^{PE}). Zafar$^{DP}_{Acc}$ is an exception as it explicitly controls the allowable accuracy drop. We hypothesize that in an effort to enforce fairness in terms of *DI*, the corresponding approaches shift the decision boundary significantly compared to LR. In contrast, approaches that target *TPRB* and *TNRB* do not need a significant boundary shift as LR's performance on these metrics is already high. The post-processing approaches, Hardt^{EO} and Pleiss^{EOP}, appear to be outliers in this observation, but as we discuss later, their accuracy drop is indicative of the poor correctness-fairness balance that is typical in post-processing. In the other three datasets, LR does not display such differences across these fairness metrics, and we do not observe significant differences in the accuracy performance of fair approaches that target demographic parity vs equalized odds.

> *Key takeaway:* Fair approaches generally trade accuracy for fairness. The compromise in accuracy is bigger when fairness-unaware approaches achieve low fairness wrt the fairness metric that a fair approach optimizes for, relative to other metrics. The tradeoff is less interpretable for correctness metrics other than accuracy, as classifiers typically do not optimize for them.

There is no single winner. All approaches succeed in improving fairness wrt the metric (and notion) they target. However, they cannot guarantee fairness wrt other notions: their performance wrt those notions is generally unpredictable. This is in line with the impossibility theorem, which states that enforcing multiple notions of fairness is impossible in the general case [20]. While we observe that approaches frequently improve on fairness metrics they do not explicitly target, this can depend on the dataset and on correlations across metrics. No approach achieves perfect fairness across all metrics. Thomas^{DP} and Thomas^{EO} come close in the German dataset, but note that this dataset contains low gender-based biases and even LR achieves reasonable fairness scores on all metrics. Note that many techniques exhibit "reverse" discrimination (the red stripes indicate discrimination against the privileged group), but these effects are generally small (a high striped bar indicates high fairness, and, thus, low discrimination in the opposite direction).

> *Key takeaway:* Approaches improve fairness on the metric they target, but their performance on other metrics is unpredictable.

Confounding factors produce different fairness assessments across metrics. Note the interesting contrast between *DI* and *CRD* on Adult (Figure 10(a)). *DI* and *CRD* essentially measure the same type of fairness, but *CRD* accounts for possible confounding effects. In Adult, LR's performance difference between *CRD* and *DI* indicates confounding factors that reduce fairness wrt *DI*. Specifically, women are strongly correlated with lower-wage occupations and fewer work hours, so when *CRD* uses occupation and working hours/week as resolving attributes, it produces high fairness scores. We observe that causal approaches, such as Zha-Wu^{PSF} and Salimi^{JF}, are particularly adept at maximizing fairness scores in *CRD* due to taking causal associations into account. Other approaches maximizing *DI* can even decrease fairness scores in *CRD* (e.g., Feld^{DP}).

> *Key takeaway:* It is important to understand the impact of confounding factors on these metrics, but we are not arguing here that *CRD* is a better metric. In fact, arguably, the fact that women are associated with low-wage occupations and low work hours may in itself be a bias we want to measure.

Pre- and in-processing approaches achieve better individual-level fairness than post-processing. Although none of the approaches in our evaluation target individual fairness explicitly, we note that pre- and in-processing tend to produce better *CD* scores than post-processing. Even for the Credit dataset (Figure 10(d)), where post-processing techniques improve the *CD* score, they do worse than pre- and in-processing on average. This is because post-processing operates on less information than pre- and in-processing, it does not assume knowledge of the attributes in the training data, and, thus, does not take similarity of individuals into account.

> *Key takeaway:* Pre- and in-processing achieve better individual-level fairness than post-processing. This is an inherent limitation of post processing, as it has no knowledge of the attributes in the training data and cannot take individual similarity into account.

Pre- and in-processing achieve better correctness-fairness balance than post-processing. Post-processing operates at a late stage of the learning process and does not have access to all the data attributes by design. As a result, it has less flexibility than pre- and in-processing. Given the fact that post-hoc correction of predictions are sub-optimal with finite training data [79], post-processing approaches typically achieve inferior correctness-fairness balance compared to other approaches. This limitation of post-processing

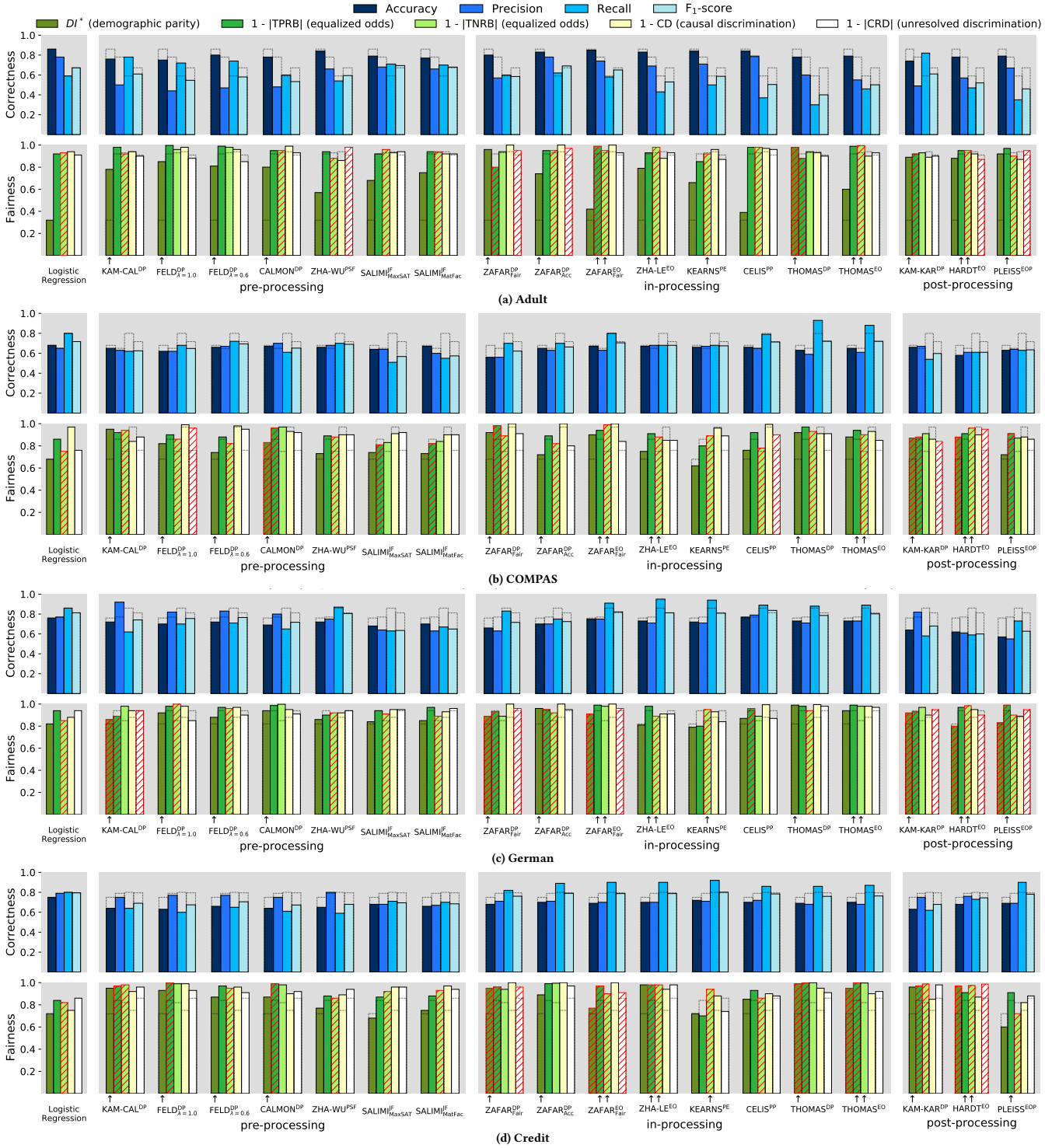


Figure 10: Correctness and fairness scores of the 18 fair classification approaches over (a) Adult, (b) COMPAS, (c) German, and (d) Credit datasets. Higher scores for correctness (fairness) metrics correspond to more correct (fair) outcomes. The bars highlighted in red denote the reverse direction of the remaining discrimination—favoring the unprivileged group more than the privileged group. The arrows (↑) denote the fairness metric(s) each approach is optimized for. The bar plots for LR are overlaid for aiding visual comparison. CALMON^{DP} failed to complete on the Credit dataset due to the large number of attributes (26); we display its performance over 22 attributes (the most it could handle).

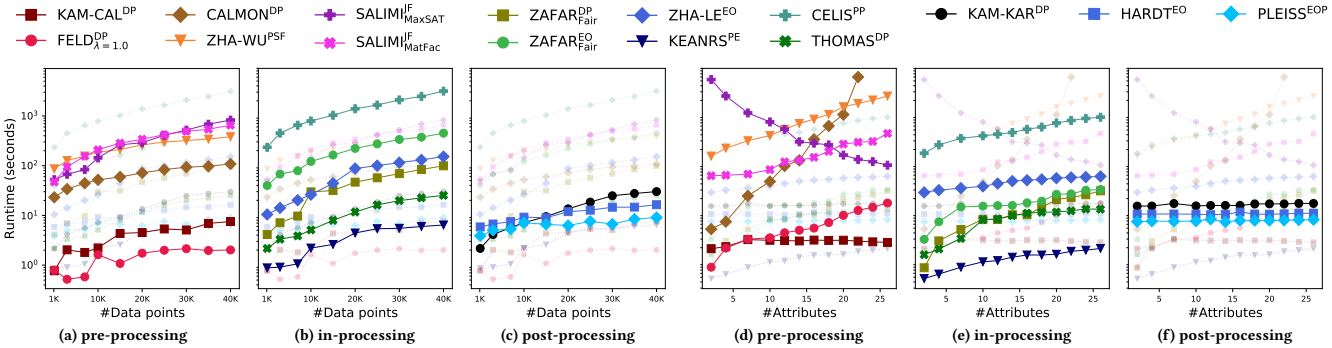


Figure 11: Results of efficiency and scalability experiments on the fair approaches. (a) – (c) show runtimes with varying data sizes in Adult dataset and (d) – (f) show runtime with varying number of attributes in Credit dataset. Note that the y-axis is in log scale.

is best highlighted in German (Figure 10(c)), where post-processing achieves on average 5-10% lower accuracy and F_1-score compared to pre- and in-processing that target the same fairness metrics. Specifically, PLEISS^{EOP} and HARDT^{EO} report the lowest accuracy and F_1-score across all techniques.

Generally, in-processing also tends to achieve 2-5% higher accuracy than pre-processing, but there is no noticeable pattern across other correctness metrics. Since in-processing modifies the training objectives directly, it has better control of the accuracy-fairness tradeoff than other methods. In contrast, there is no direct mapping between the extent of repair during pre-processing and the compromise in accuracy, so pre-processing approaches cannot directly control this tradeoff. However, we cannot conclude that in-processing is always better at balancing correctness and fairness. Rather, pre-processing approaches require appropriate tuning of the level of repair to achieve the desired correctness-fairness balance, and further investigation needs to be done in this regard.

Key takeaway: Pre- and in-processing achieve better correctness and fairness compared to post-processing. In-processing handles the accuracy-fairness tradeoff most effectively, but pre-processing can see gains from proper tuning of the level of repair.

4.3 Efficiency and Scalability

In this section, we study the runtime behavior of all approaches. We do not present separate variants of the same approach unless they differ significantly in behavior. We compute the total runtime of each approach as pre-processing time (if any) + training time + post-processing time (if any). We subtract from all methods the runtime of LR, so that what we report is the overhead each approach introduces over the fairness-unaware method.

Our first experiment investigates the efficiency and scalability of the fair approaches as the number of data points increases. We used the Adult dataset, as it contains the highest number of data points, and executed a new instance of each approach with a different number of data points (from 1K to 40K) sampled from the dataset. Our second experiment explores the runtime behavior of approaches as the number of attributes increases. We used the Credit dataset, as it contains the highest number of attributes, and executed a new instance of each approach with a different number of attributes (from 2 to 26). We present the results in Figure 11.

Post-processing approaches are generally most efficient and scalable. Post-processing approaches tend to be very efficient, as their mechanisms are less complex compared to pre- and in-processing approaches. As a result, they scale well wrt increasing data sizes and they are not affected by increase in the number of attributes. A few pre- and in-processing techniques like KAM-CAL^{DP} and KEARNS^{PE} do perform better than post-processing, but this does not hold for most other techniques in their categories.

Key takeaway: Post-processing approaches are more efficient and scalable than pre- and in-processing approaches. Pre- and in-processing approaches generally incur higher runtimes, which depend on their computational complexities.

Causal computations incur sharp runtime penalties. An important observation from Figure 11(a) is that causal mechanisms—such as ZHA-WU^{PSF} and SALIMI^{JF}—have significantly higher runtimes compared to other pre-processing approaches. In fact, both variations of SALIMI^{JF} are NP-hard in nature. Simply, discovering causal associations from data is more complex than non-causal associations. CALMON^{DP} also demonstrates high runtimes, in its case due to relying on solving convex optimization problems, and very poor scalability with increasing attributes (Figure 11(d)).

Key takeaway: Causality-based mechanisms incur higher runtimes. Other complex mechanisms also lead to efficiency and scalability challenges.

Pre-processing approaches scale well with increasing data sizes, but tend to scale poorly with increasing number of attributes. As we noted, there is a clear separation between the inherently more complex pre-processing methods (ZHA-WU^{PSF}, SALIMI^{JF}, and CALMON^{DP}) and the rest (KAM-CAL^{DP} and FELD^{DP}). In fact, KAM-CAL^{DP} and FELD^{DP} perform on par with or better than post-processing techniques in terms of efficiency, and generally better than most in-processing approaches. Generally, pre-processing demonstrates more robust scaling behavior wrt the data size than the number of attributes. In fact, the runtime of several pre-processing approaches appears to grow exponentially with the number of attributes (Figure 11(d)). CALMON^{DP} did not converge for more than 22 attributes as its complexity is tied to the number of attributes. Causality-based approaches display similar challenges. There is, however, an interesting contrast between SALIMI$^{JF}_{MaxSAT}$ and SALIMI$^{JF}_{MatFac}$. The number of constraints in SALIMI$^{JF}_{MaxSAT}$ increases rapidly with fewer

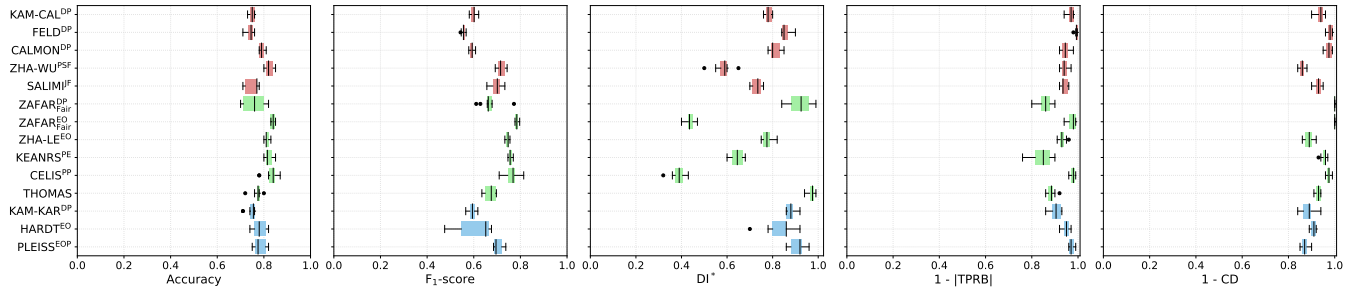


Figure 12: Variance of the fair approaches in terms of accuracy, F_1-score, DI, TPRB, and CD on arbitrary folds over the Adult dataset.

attributes, resulting in higher runtimes. In contrast to other techniques, its performance improves as the number of attributes grows.

In-processing approaches scale better than pre-processing ones wrt the number of attributes, but are more affected by the data size. In-processing techniques show a slightly sharper rise in runtime when the data size increases compared to pre-processing approaches (Figure 11(b)). However, in-processing techniques scale more gracefully than pre-processing ones with the number of attributes. Their runtime does increase, since the higher number of attributes increases the complexity of the decision boundary in optimization problems, but it is generally lower than pre-processing, which typically performs data modification on a per-attribute basis.

> *Key takeaway:* Pre-processing approaches are generally more affected by the number of attributes than the data size. In-processing approaches are generally more affected by the data size than the number of attributes.

4.4 Stability

We evaluate the stability of all the approaches through a variance test on their correctness and fairness. We executed each fair approach 10 times with random folds, using 66.67% of the data for training and the rest for testing. We report our findings on the stability of two correctness metrics (accuracy and F_1-score) and three fairness metrics (*DI*, *TPRB*, and *CD*) over Adult (Figure 12); the results are similar for other accuracy and fairness metrics, and over the other datasets (full results are in the Appendix).

Approaches are generally stable. Most approaches show low variance and have a very small number of outliers. Although it exhibits low variance in accuracy, HARDT^{EO} has the highest variance in F_1-score. ZAFAR$^{DP}_{FAIR}$ shows slightly high variance in accuracy and *DI*, but is stable on the other metrics. In general, post-processing approaches exhibit slightly high variance in accuracy, F_1-score, and *DI*.

> *Key takeaway:* All approaches generally exhibit low variance in terms of correctness and fairness over different train-test splits. High-variance behavior is rare, and there is no significant trend across the dimension of pre-, in-, and post-processing.

5 LESSONS AND DISCUSSION

The goal of our work has been to bring some clarity to the vast and diverse landscape of fair classification research. Work on this topic has spanned multiple disciplines with different priorities and focus, resulting in a wide range of approaches and diverging evaluation goals. Data management research has started making important contributions to this area, and we believe that there are a lot of opportunities for impact and synergy. Through our evaluation, we aimed in particular to identify areas and opportunities where data management contributions appear better-suited to be successful. We discuss these general guidelines here.

Pre-processing approaches are a natural fit but exhibit scalability challenges. Data management research has primarily focused on the pre-processing stage, as data manipulations create a natural fit. However, our evaluation showed that pre-processing methods tend to not scale robustly with the number of attributes. Research in pre-processing methods should be mindful of problem settings where the high data dimensionality may lead to a poor fit. However, this observation also points to an opportunity that plays squarely into the strengths of the data management community, as efforts can focus directly on attacking this scalability challenge. Already, data management researchers made contributions in this direction (e.g., SALIMI^{JF} has a parallel implementation, which was not suitable for our evaluation as other approaches are single-threaded), and improvements here are likely to lead to more impact.

Similarly, we noted that in-processing techniques generally outperform others in handling the correctness-fairness tradeoff directly. However, pre-processing methods have the potential to improve this balance through appropriate tuning of the data repair levels, and further investigation can help in that regard.

Finally, causality-based approaches produce sophisticated repairs, but impose a significant runtime penalty. KAM-CAL^{DP} and FELD^{DP} use simpler repairs, resulting in orders of magnitude better runtime performance, while they maintain competitive correctness-performance tradeoff as well.

Applicability of pre-processing. Pre-processing possesses the flexibility of being model agnostic, and can be used when access and modifications to the model are not possible. However, there can also be practical constraints to modifying training data, as this may violate anti-discrimination laws [5]. Additionally, pre-processing repairs data on the assumption that model predictions will follow the ground truth. But, it cannot enforce fairness notions that target the correctness of predictions across sensitive groups, as it cannot make assumptions on the correctness of predictions after the learning step. This means that metrics such as equalized odds and predictive parity cannot be easily handled in the pre-processing stage. As we saw in our evaluation, fairness as measured by different metrics can diverge, and it is important to follow the application requirements before attacking a problem setting with pre-processing methods.

Synergy with ML research. Our analysis noted that in-processing techniques exhibit better control of the correctness-fairness tradeoff and may be hard to beat in that regard. However, their performance scales worse with increasing data size compared to pre-processing approaches. Generally, runtime performance is often overlooked in machine learning research, and data management contributions can likely have impact in improving in-processing approaches in that regard.

We hope that our analysis will be helpful to outline useful perspectives and directions to data management research in fair classification. To the best of our knowledge, ours is the broadest evaluation and analysis of work in this area, and can contribute to a useful roadmap for the research community.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML*.

[2] Saba Ahmadi, Sainyam Galhotra, Barna Saha, and Roy Schwartz. 2020. Fair Correlation Clustering. *CoRR* abs/2002.03508 (2020). arXiv:2002.03508 https://arxiv.org/abs/2002.03508

[3] Abolfazl Asudeh and HV Jagadish. 2020. Fairly evaluating and scoring items in a data set. *Proceedings of the VLDB Endowment* 13, 12 (2020).

[4] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*. 1259–1276.

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

[6] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[7] Vidmantas Bentkus et al. 2004. On Hoeffding's inequalities. *The Annals of Probability* 32, 2 (2004), 1650–1673.

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

[9] Brian Borchers and Judith Furman. 1998. A two-phase exact algorithm for MAX-SAT and weighted MAX-SAT problems. *Journal of Combinatorial Optimization* 2, 4 (1998), 299–306.

[10] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[11] Larry Brown. 1967. The conditional level of Student's t test. *The Annals of Mathematical Statistics* 38, 4 (1967), 1068–1071.

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

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

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

[15] Simon Caton and Christian Haas. 2020. Fairness in Machine Learning: A Survey. *arXiv preprint arXiv:2010.04053* (2020).

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

[17] Shunqin Chen, Zhengfeng Guo, and Xinlei Zhao. 2020. Predicting Mortgage Early Delinquency with Machine Learning Methods. *European Journal of Operational Research* (2020).

[18] Yuh-Wen Chen and Moussa Larbani. 2006. Two-person zero-sum game approach for fuzzy multiple attribute decision making problems. *Fuzzy Sets and Systems* 157, 1 (2006), 34–51.

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

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

[21] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020).

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

[23] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* (2016).

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

[25] Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. 2012. Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*. Citeseer.

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

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

[28] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation* 80 (2016), 38.

[29] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. 1918–1921.

[30] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[31] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. 498–510.

[32] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R. Varshney. 2020. Fair Data Integration. *CoRR* abs/2006.06053 (2020). arXiv:2006.06053 https://arxiv.org/abs/2006.06053

[33] German Credit Risk 2020. German Credit Risk- Kaggle. https://www.kaggle.com/uciml/german-credit.

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

[35] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1809–1818.

[36] William W Hager and Sanjoy K Mitter. 1976. Lagrange duality theory for convex control problems. *SIAM Journal on Control and Optimization* 14, 5 (1976), 843–856.

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

[38] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20 (2011), 217–240.

[39] Gareth P Jones, James M Hickey, Pietro G Di Stefano, Charanpal Dhanjal, Laura C Stoddart, and Vlasios Vasileiou. 2020. Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986* (2020).

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

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

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

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

[44] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2564–2572.

[45] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[46] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[47] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*. 4842–4852.

[48] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*.

[49] Ronny Kohavi and Barry Becker. 1994. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Adult

[50] Caitlin Kuhlman and Elke Rundensteiner. 2020. Rank aggregation algorithms for fair consensus. *Proceedings of the VLDB Endowment* 13, 12 (2020).

[51] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*. 4066–4076.

[52] Vincent Labatut and Hocine Cherifi. 2012. Accuracy measures for the comparison of classifiers. *arXiv preprint arXiv:1207.3790* (2012).

[53] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems* 33 (2020).

[54] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment* 13, 4 (2019), 506–518.

[55] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).

[56] Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*. 556–562.

[57] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *stat* 1050 (2015), 3.

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

[59] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[60] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[61] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, Vol. 1170.

[62] Alejandro Noriega-Campero, Michiel A Bakker, Bernardo Garcia-Bulle, and Alex'Sandy' Pentland. 2019. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 77–83.
[63] Judea Pearl. 2009. *Causality*. Cambridge university press.
[64] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.
[65] Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems*. 677–688.
[66] Bilal Qureshi, Faisal Kamiran, Asim Karim, Salvatore Ruggieri, and Dino Pedreschi. 2019. Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems* (2019), 1–13.
[67] Jonathan Rothwell. 2014. How the war on drugs damages black social mobility. *The Brookings Institution, published Sept* 30 (2014).
[68] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. 6414–6423.
[69] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*. 793–810.
[70] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair pca: One extra dimension. In *Advances in Neural Information Processing Systems*. 10976–10987.
[71] Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. 1998. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research* 33, 1 (1998), 65–117.
[72] Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. 2016. Disciplined convex-concave programming. In *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 1009–1014.
[73] Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*.
[74] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004.
[75] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel J Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2015. Discovering unwarranted associations in data-driven applications with the fairtest testing toolkit. *CoRR, abs/1510.02377* (2015).
[76] Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani. 2012. Websites vary prices, deals based on users' information. *Wall Street Journal* 10 (2012), 60–68.
[77] Vladimir N Vapnik and A Ya Chervonenkis. 2015. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*. Springer, 11–30.
[78] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
[79] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*. 1920–1953.
[80] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*. 3404–3414.
[81] An Yan and Bill Howe. 2019. Fairst: Equitable spatial and temporal demand prediction for new mobility systems. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 552–555.
[82] I Cheng Yeh. 2016. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[83] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.
[84] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*. 229–239.
[85] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.
[86] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.
[87] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[88] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3929–3935.

A DESCRIPTION OF FAIR APPROACHES

In this section, we provide detailed discussion of the fair approaches that we evaluate in this paper.

A.1 Pre-processing Approaches

A.1.1 **KAM-CAL.** Kamiran and Calders [40] introduce a pre-processing approach that targets the notion of demographic parity. We refer to this approach as KAM-CAL. Assuming that the predictions $\hat{Y}$ reasonably approximates the ground truth Y, KAM-CAL argues that $\hat{Y}$ is likely to be independent of the sensitive attribute S, when the classifier is deployed, if Y and S are independent in the training data. To this end, KAM-CAL samples tuples from the training dataset $\mathcal{D}$ to create a modified training dataset $\mathcal{D}'$ in a way that ensures that Y and S are independent in $\mathcal{D}'$. This is based on the intuition that the classifier is likely to learn the independence from $\mathcal{D}'$ and will ensure demographic parity when deployed.

If S and Y are independent in $\mathcal{D}$, then $\forall s \in S$ and $\forall y \in Y$, their expected joint probability $Pr_{exp}(S = s \wedge Y = y)$ should be sufficiently close to their observed joint probability $Pr_{obs}(S = s \wedge Y = y)$. These probabilities (over $\mathcal{D}$) are computed using the following formulas:

$$Pr_{exp}(S = s \wedge Y = y) := \frac{|\{t : S_t = s\}|}{|\mathcal{D}|} \cdot \frac{|\{t : Y_t = y\}|}{|\mathcal{D}|}$$

$$Pr_{obs}(S = s \wedge Y = y) := \frac{|\{t : S_t = s \wedge Y_t = y\}|}{|\mathcal{D}|}$$

If Pr_{obs} is different from Pr_{exp}, then S and Y are not independent in $\mathcal{D}$. KAM-CAL's goal is to modify $\mathcal{D}$ to obtain $\mathcal{D}'$ such that the differences between the expected and the observed probabilities are mitigated. To achieve this, KAM-CAL employs a weighted sampling technique that compensates for the differences in Pr_{exp} and Pr_{obs}. The technique involves computing a weight for each tuple in $\mathcal{D}$ and then sampling the tuples from $\mathcal{D}$, with probability proportional to their weights, to construct $\mathcal{D}'$. The weight $w(t)$ of a tuple $t \in \mathcal{D}$ is computed as:

$$w(t) = \frac{Pr_{exp}(S = S_t \wedge Y = Y_t)}{Pr_{obs}(S = S_t \wedge Y = Y_t)}$$

This weighting scheme guarantees that Pr_{exp} and Pr_{obs} are sufficiently close over $\mathcal{D}'$, which implies that Y and S are independent in $\mathcal{D}'$. KAM-CAL also provides empirical evidence that classifiers trained on $\mathcal{D}'$ indeed satisfy demographic parity.

Implementation. We collected the source code for KAM-CAL from the open source AI Fairness 360 library.[2]

A.1.2 **FELD.** Feldman et al. [26] propose a pre-processing approach that also enforces demographic parity. We refer to this approach as FELD. FELD argues that demographic parity can be ensured if the marginal distribution of each $X \in \mathbb{X}$ is similar across the privileged and the unprivileged groups in the training data. The basis of their argument is that if a model learns from such data, it is likely to predict based on attributes that are independent of S, which in turn will satisfy demographic parity within the model's predictions. Unlike KAM-CAL, which does not modify attribute values, FELD directly modifies the values for each attribute X.

Given the training data $\mathcal{D} = [\mathcal{D}_{\mathbb{X}}, \mathcal{D}_S; \mathcal{D}_Y]$ with the schema $(\mathbb{X}, S; Y)$, FELD produces a modified dataset $\mathcal{D}' = [\mathcal{D}'_{\mathbb{X}}, \mathcal{D}_S; \mathcal{D}_Y]$ where the marginal distribution of each attribute is similar across the privileged and the unprivileged groups. FELD repairs values of each individual attribute separately to equalize the marginal distribution of the sensitive groups for each attribute. To this end, FELD determines the quantile of each value $x \in \mathcal{D}_X$ and replaces x with the median of the corresponding quantiles from the original marginal distributions $Pr(\mathcal{D}_X \mid S = 1)$ and $Pr(\mathcal{D}_X \mid S = 0)$. This repair produces the modified attribute $\mathcal{D}'_X$ such that $Pr(\mathcal{D}'_X \mid S = 1) = Pr(\mathcal{D}'_X \mid S = 0)$, and, thus, ensures that the modified attribute is independent of the sensitive attribute.

[2] https://github.com/Trusted-AI/AIF360/tree/master/aif360/algorithms/preprocessing

Repeating the repair process for all attributes produces the modified $\mathcal{D}'_{\mathbb{X}}$ and the modified dataset $\mathcal{D}'$. The level of repair is controlled through a hyper-parameter $\lambda \in [0, 1]$, where $\lambda = 0$ yields the unmodified dataset and $\lambda = 1$ implies that the values within each attribute are completely moved to the median.

Implementation. We collected the source code for FELD from the AI Fairness 360 library.[2] As the preferred value of λ is application-specific, we choose two values (1.0 and 0.6) in our evaluation to highlight its impact on the performance.

A.1.3 CALMON. Calmon et al. [14] propose a pre-processing approach that also enforces demographic parity. We refer to this approach as CALMON. Given the joint distribution associated with the training data $\mathcal{D}$, CALMON computes a new distribution to transform $\mathbb{X}$ and Y such that the dependency between Y and S is reduced, without significantly distorting the data distribution. The new joint distribution yields repaired training data $\mathcal{D}' = [\mathcal{D}'_{\mathbb{X}}, \mathcal{D}_S; \mathcal{D}'_Y]$.

To compute the new distribution, CALMON constructs the following constraints that must be satisfied: (1) the difference between $Pr(\mathcal{D}'_Y \mid S = 0)$ and $Pr(\mathcal{D}'_Y \mid S = 1)$ is below an allowable threshold, (2) the new joint distribution is sufficiently close to the original one, and (3) no attribute value in $\mathcal{D}_{\mathbb{X}}$ is substantially distorted to compute $\mathcal{D}'_{\mathbb{X}}$. CALMON then formulates a convex optimization problem that searches for the optimal new distribution subject to the constraints. The resulting new distribution maps each tuple from $\mathcal{D}$ to the modified dataset $\mathcal{D}'$ and classifiers learned on $\mathcal{D}'$ is expected to satisfy demographic parity.

Implementation. We collected the source code for CALMON from the AI Fairness 360 library.[2] Further, as CALMON could not operate on more than 22 attributes on the Credit dataset over our system, we dropped 4 attributes with the least information gain.

A.1.4 ZHA-WU. Zhang, Wu, and Wu [88] propose a pre-processing approach that targets *path-specific fairness*: a causal notion that uses graphical causal models to ensure that causal effects of the sensitive attribute are not carried to the prediction through any direct or indirect path, i.e., S does not have any causal association with $\hat{Y}$. We refer to this approach as ZHA-WU. Using the training data $\mathcal{D} = [\mathcal{D}_{\mathbb{X}}, \mathcal{D}_S; \mathcal{D}_Y]$, ZHA-WU constructs a graphical causal model to estimate the effect of intervening on $\mathcal{D}_S$ and determines its causal association with $\mathcal{D}_Y$. Then they repair $\mathcal{D}_Y$ minimally to produce $\mathcal{D}'_Y$ such that all causal associations between $\mathcal{D}'_Y$ and $\mathcal{D}_S$ are removed. Then the classifiers trained on the modified training data $\mathcal{D}' = [\mathcal{D}_{\mathbb{X}}, \mathcal{D}_S; \mathcal{D}'_Y]$ are expected to satisfy path-specific fairness, under the assumption that the distribution of the predictions made by a classifier follows the distribution of the ground truth in the training data.

To repair $\mathcal{D}_Y$, ZHA-WU first verifies if $\mathcal{D}_Y$ violates path-specific fairness. Specifically, $\mathcal{D}_S$ is a direct or indirect cause of $\mathcal{D}_Y$ if intervening on $\mathcal{D}_S$ changes the expectations of $\mathcal{D}_Y$. ZHA-WU utilizes $\mathcal{D}$ to construct the graphical causal model and estimates the effect of intervening on $\mathcal{D}_S$ as the expected difference in $\mathcal{D}_Y$ when $\mathcal{D}_S$ changes from privileged to unprivileged. Instead of measuring causal association through all paths between $\mathcal{D}_S$ and $\mathcal{D}_Y$ in the causal graph, ZHA-WU can measure this association through specific paths if desired. Path-specific fairness is violated if the expected difference in $\mathcal{D}_Y$ is above some threshold ϵ.

Next, ZHA-WU designs an optimization problem to repair $\mathcal{D}_Y$ such that its direct and indirect causal associations with $\mathcal{D}_S$ are removed, and the causal model is minimally altered. The modified training dataset $\mathcal{D}'$ is then used to train classifiers that enforce path-specific fairness. Note that an accurate representation of the causal model depends on the training data, and ZHA-WU allows alternative causal models that can be constructed with domain knowledge.

Implementation. We retrieved the source code for ZHA-WU from the authors' website.[3] In accordance with the original paper, we construct the causal networks using the open source software TETRAD [71] and set the value for ϵ to 0.05.

A.1.5 SALIMI. Salimi et al. [69] propose a pre-processing approach that enforces *justifiable fairness*: a causal fairness notion that prohibits causal dependency between the sensitive attribute S and the prediction $\hat{Y}$, except through admissible attributes. We refer to this approach as SALIMI. Unlike other causal mechanisms, SALIMI does not require access to the causal model. SALIMI assumes that $\hat{Y}$ is likely to be fair if a classifier is trained on data $\mathcal{D}$ where ground truth Y satisfies the target fairness notion. To that end, it expresses justifiable fairness as an integrity constraint and repairs $\mathcal{D}$ to ensure that the constraint holds on the repaired training data $\mathcal{D}'$. Unlike KAM-CAL, SALIMI does not modify the attributes and only repairs $\mathcal{D}$ by inserting or deleting tuples.

As SALIMI does not depend on the causal model, it translates the condition for justifiable fairness into an integrity constraint that must hold over the training data. SALIMI partitions all attributes, except the ground truth, into two disjoint sets: *admissible* (**A**) and *inadmissible* (**I**). **A** contains the attributes that are allowed to influence or have causal associations with prediction $\hat{Y}$, while **I** contains the rest of the attributes. Given **A** and **I**, justifiable fairness holds in $\mathcal{D}$ if Y is independent of **I** conditioned on **A**. If the probability distribution associated with $\mathcal{D}$ is uniform,[4] this integrity constraint can be checked through the following multi-valued dependency: $\mathcal{D} = \Pi_{\mathbf{A}Y}(\mathcal{D}) \bowtie \Pi_{Y\mathbf{I}}(\mathcal{D})$.

The goal of SALIMI is then to minimally repair $\mathcal{D}$ to form a new training dataset $\mathcal{D}'$, such that the multi-valued dependency is satisfied. SALIMI leverages techniques from maximum satisfiability [9] and matrix factorization [56] to compute the minimal repair of $\mathcal{D}$ that produces the optimal $\mathcal{D}'$ for training classifiers. However, these techniques are NP-hard and application-specific knowledge is generally needed to determine the sets of admissible and inadmissible attributes.

Implementation. We collected the source code for SALIMI from the authors via email, as no public repository is available. Following the original paper, we choose race, gender, marital/relationship status as inadmissible attributes whenever applicable, and the rest of the attributes as admissible. Moreover, Salimi et al. discuss a second variation of $\text{SALIMI}^{\text{IF}}_{\text{MaxSAT}}$ that partially repairs the data, but we do not include it as there are no instructions on how to tune the level of repair for that. Lastly, although there are experiments in the original paper that discuss techniques to partition the training data and repair them in parallel, our evaluation is limited to a single-threaded implementation.

[3] https://www.yongkaiwu.com/publication/zhang-2017-causal/

[4] Datasets do not always have uniform probability distribution in practice and additional pre-processing is required to ensure that.

A.2 In-processing Approaches

Zafar. Zafar et al. [83, 85] propose two in-processing approaches to enforce demographic parity and equalized odds. We refer to them as Zafar$^{\text{DP}}$ and Zafar$^{\text{EO}}_{\text{FAIR}}$, respectively. Both of these approaches translate their corresponding fairness notion to a convex function of the classifier parameters, and compute the optimal parameters that minimize prediction errors while satisfying the notion.

To compute the optimal fair classifier, Zafar first formulates the learning process as a constrained optimization problem. Given the training data $\mathcal{D}$, the task of a classifier is to learn a decision boundary that separates the tuples according to the ground truth. The optimal decision boundary, defined by a set of parameters θ, is the one that minimizes a convex loss function $L(\theta)$ that measures the cost of prediction errors. For any tuple t, the signed distance from the decision boundary determines the prediction. Specifically, $\hat{Y}_t = 1$ if $d_\theta(\mathbb{X}_t) \geq 0$, where $d_\theta(\mathbb{X}_t)$ denotes the signed distance. Zafar does not explicitly use S to determine the prediction, rather they utilize S to define the fairness constraint only.

Zafar$^{\text{DP}}$ introduces a proxy constraint for demographic parity, as directly including the notion as a constraint leads to non-convexity in the loss function.[5] Zafar$^{\text{DP}}$ utilizes d_θ as a proxy for $\hat{Y}$ and argues that the empirical covariance between the sensitive attribute and the signed distance from the decision boundary is approximately zero, if the prediction of a classifier is independent of the sensitive attribute. As covariance is a convex function of θ, it can be used define the proxy constraint for demographic parity. Formally, covariance is computed as: $cov = \frac{1}{|\mathcal{D}|}\sum_{t\in\mathcal{D}}(S_t - \bar{S})d_\theta(\mathbb{X}_t)$, where $\bar{S}$ denotes the mean of S. Given the proxy constraint, Zafar$^{\text{DP}}$ proposes the following two variations that work under different constraint settings:

- **Maximizing accuracy under fairness constraint.** This variation (Zafar$^{\text{DP}}_{\text{FAIR}}$) computes the optimal classifier by minimizing $L(\theta)$ under the condition that $cov \approx 0$.
- **Maximizing fairness under accuracy constraint.** This variation (Zafar$^{\text{DP}}_{\text{ACC}}$) minimizes cov as much as possible while ensuring $L(\theta)$ is below a specified threshold. This is to avoid cases where enforcing $cov \approx 0$ leads to high loss in the first variation.

Both of the above variations produce a fair classifier that approximately satisfies demographic parity. Similar to Zafar$^{\text{DP}}$, Zafar$^{\text{EO}}_{\text{FAIR}}$ introduces a proxy constraint for equalized odds. In particular, Zafar$^{\text{EO}}_{\text{FAIR}}$ proposes to use the covariance between S and d_θ of the misclassified tuples, since covariance is approximately zero when a classifier satisfies equalized odds. This covariance is computed as: $cov = \frac{1}{|\mathcal{D}|}\sum_{t\in\mathcal{D}}(S_t - \bar{S})g_\theta(\mathbb{X}_t)$, where $g_\theta(\mathbb{X}_t) = -d_\theta(\mathbb{X}_t)$ if tuple t is misclassified, and 0 otherwise. While this proxy is still not a convex function of θ, Zafar$^{\text{EO}}_{\text{FAIR}}$ efficiently computes classifier parameters that maximize prediction accuracy under this proxy constraint through a disciplined convex-concave program [72].

Implementation. We collected the source code for Zafar from the authors' public repository.[6] We set all the hyper-parameters following the instructions specified within the source code (more details are in the authors' repository).

Zha-Le. Zhang, Lemoine, and others [87] propose an in-processing approach that can enforce demographic parity, equalized odds, or equal opportunity, by leveraging *adversarial learning*, a technique where a classifier and an adversary with mutually competing goals are trained together. We refer to this approach as Zha-Le. Given the training data $\mathcal{D} = (\mathbb{X}, S; Y)$, the goal of a classifier f is to maximize the accuracy of prediction $\hat{Y}$, while an adversary a attempts to correctly predict the sensitive attribute using $\hat{Y}$ (and Y). Zha-Le enforces the target notion of fairness by designing the classifier to converge to optimal parameters such that $\hat{Y}$ does not contain any information about S that the adversary can exploit.

In order to determine the optimal parameters, classifier f minimizes a loss function $L_f(\hat{Y}, Y)$. Adversary a receives both $\hat{Y}$ and Y if equalized odds or equal opportunity is the target notion, otherwise a only has access to $\hat{Y}$ if demographic parity is enforced. The loss of adversary is denoted as $L_a(\hat{S}, S)$. Both the classifier and adversary apply gradient based optimizations [10] to iteratively update their parameters. Adversary a updates its parameters in a direction that minimizes L_f, while the classifier f only updates its parameters in a direction that both decreases L_f and increases L_a. This process of update guarantees that f converges to a solution where $L_f(\hat{Y}, Y)$ is minimized while $L_a(\hat{S}, S)$ is approximately equal to the entropy of S, i.e., adversary gains no information about S from $\hat{Y}$ (and Y). Hence, the optimal classifier satisfies the target fairness notion.

Implementation. We collected the source code for Zha-Le from the open source AI Fairness 360 library.[7]

Kearns. Kearns et al. [44] propose an in-processing approach that enforces demographic parity and predictive equality, a notion that requires equal *FPR* for the privileged and the unprivileged groups. We refer to this approach as Kearns. Kearns approximately enforces the target fairness notion within a large set of subgroups[8] defined using one or more sensitive attributes (or user-specified attributes). To that end, Kearns solves a constrained optimization problem to obtain optimal classifier parameters such that the proportion of positive outcomes (demographic parity) or FPR (predictive equality) is approximately equal to that of the population.

Kearns begins by formulating the learning process and constraint for the target fairness notion. Let $f : f(\mathbb{X}, S) \rightarrow \hat{Y}$ be a classifier learned over training data $\mathcal{D} = (\mathbb{X}, S, Y)$. Moreover, let G be the set of the subgroups for which fairness must be ensured. Each $g \in G$ indicates a subgroup such that $g(\mathbb{X}_t, S_t) = 1$ means tuple t belongs to subgroup g. If predictive equality is the target notion, a group function $\beta(g) = Pr(\hat{Y} = 1 \mid Y = 0) - Pr(\hat{Y} = 1 \mid Y = 0, g(\mathbb{X}, S) = 1)$ denotes the difference between overall *FPR* and *FPR* for group g. The fairness constraint is formally expressed as: $\alpha(g)\beta(g) \leq \gamma, \forall g \in G$, where $\alpha(g)$ denotes the proportion of tuples in group g in order to exclude very small groups from calculation and γ is a tolerance parameter. Similar $\alpha(g)$ and $\beta(g)$ can be derived for demographic parity.

Next, Kearns constructs the following optimization problem to compute optimal f that minimizes a loss function $l(\hat{Y}, Y)$:

[5] Non-convex functions are computationally harder to optimize than convex functions.
[6] https://github.com/mbilalzafar/fair-classification
[7] https://github.com/Trusted-AI/AIF360/tree/master/aif360/algorithms/inprocessing
[8] The number of subgroups must be bounded by the classifier's VC dimension [77].

$$\min_f \mathbb{E}[l(\hat{Y}, Y)]$$
$$s.t.\ \alpha(g)\beta(g) \leq \gamma,\ \forall g \in G$$

While this optimization problem can be computationally hard in the worst case, Kearns computes an approximate solution by solving an equivalent zero-sum game [18] in polynomial time and the optimal classifier approximately satisfies the target fairness notion.

Implementation. We collected the source code for Kearns from the open source AI Fairness 360 library.[7] The current version does not include any implementation for demographic parity, and, thus, our evaluation is limited to predictive equality. We use $\gamma = 0.005$, as suggested in the source code.

Celis. Celis et al. [16] propose an in-processing approach that supports multiple fairness notion within a single framework. We refer to this approach as Celis. Celis can accommodate a wide range of notions: predictive parity, demographic parity, equalized odds, and conditional accuracy equality. Celis reduces each fairness notion to a linear function and presents an approach to solve the resulting linear constrained optimization problem for obtaining a fair classifier that minimizes prediction error.

In order to derive the fairness constraint, Celis first partitions the training data $\mathcal{D} = (\mathbb{X}, S, Y)$ into groups according to the sensitive attribute. Let G be the set of groups and each $g_i \in G$ denotes a group such that $g_i = (\mathbb{X}, S = i, Y) \subseteq \mathcal{D}$. For each group in G, Celis then defines $q_i(f)$ that is a linear function or quotient of linear functions of $Pr(\hat{Y} = 1 \mid g_i, \varepsilon_i)$, where ε_i can be any event relevant to the target fairness notion. Intuitively, $q_i(f)$ represents the performance of classifier f for group g_i. For example, $q_i(f)$ represents the probability of positive outcome when the target notion is demographic parity. Given the function, a fairness notion can be expressed as the following constraint: $\frac{\min_{i \in S}\ q_i(f)}{\max_{i \in S} q_i(f)} \geq \tau$, where $\tau \in [0, 1]$ denotes a tolerance parameter. $\tau = 1$ implies that a classifier's performance must be equal across all groups. Multiple constraints can be derived similarly if multiple notions need to be enforced simultaneously.

Given the fairness constraint, Celis then formulates the process of finding the optimal f as the following constrained optimization problem:

$$\min_f Pr(f(\mathbb{X}) \neq Y)$$
$$s.t.\ \frac{\min_{i \in S}\ q_i(f)}{\max_{i \in S} q_i(f)} \geq \tau$$

To solve the above problem efficiently, Celis solves its dual instead using Lagrange duality [36], which produces an approximately fair classifier. This fair classifier can only guarantee $\min_{i \in S} q_i(f) \geq \tau \cdot \max_{i \in S}\ q_i(f) - \epsilon - k$, where $\epsilon > 0$ represents some error that results from the approximation and k denotes additional error from estimating the probability distribution of data from samples in $\mathcal{D}$.

Implementation. We collected the source code for Celis from the open source AI Fairness 360 library.[7] We use $\tau = 0.8$ as suggested in the source code. Further, we noted that the difference in accuracy was minimal ($\leq 1\%$) for any $\tau \in [0.8, 1.0]$, and, thus, further hyper-parameter tuning was not necessary.

Thomas. Thomas et al. [74] propose an in-processing approach that can enforce demographic parity, equalized odds, equal opportunity, and predictive equality. We refer to this approach as Thomas. Given a training data $\mathcal{D}$ and a target fairness notion, Thomas ensures that a classifier f trained on $\mathcal{D}$ only picks solutions that satisfy the fairness notion with high probability. Thomas computes an upper bound (with high confidence) of the maximum possible fairness violation that a classifier can incur at test time, and returns optimal classifier parameters for which this worst possible violation is within an allowable threshold.

Given a function g that quantifies discrimination according to the target fairness notion and an objective function L denoting a classifier's correctness, Thomas's goal is formalized below:

$$\operatorname{argmax}_f L(f)$$
$$s.t.\ Pr(g(f(\mathcal{D})) \leq 0) \geq 1 - \delta,$$

Here, $1 - \delta$ denotes the confidence upper bound. While Thomas allows multiple g to specify multiple fairness constraints simultaneously, it fails to compute a feasible solution if the specified fairness notions cannot be enforced at the same time. In order to compute the optimal fair solution, Thomas splits the training data into two partitions: $\mathcal{D}_1$ and $\mathcal{D}_2$. Thomas then uses gradient descent to compute a candidate solution that maximizes the objective function on $\mathcal{D}_1$. Using $\mathcal{D}_2$, Thomas derives an upper bound on the amount of discrimination that the candidate solution can incur. This upper bound is computed using concentration inequalities, such as Hoeffding's inequality [7] or Student's t-test [11]; and denotes the maximum amount of discrimination that can occur, with a confidence of $1 - \delta$. Finally, Thomas selects the candidate solution as the optimal solution if the upper bound is acceptable in the context of the problem, and returns no solution otherwise.

Implementation. We collected the source code for Thomas from the authors via email, as no public repository is available. Although Thomas supports multiple notions of fairness (Figure 8), we exclude two notions—equal opportunity and predictive equality—from our evaluation, as equalized odds encompasses both these notions. We use $\delta = 0.05$, in accordance with the paper.

A.3 Post-processing Approaches

A.3.1 Kam-Kar. Kamiran, Karim, and others [42] propose a post-processing approach that enforces demographic parity. We refer to this approach as Kam-Kar. Kam-Kar is based on the intuition that discriminatory decisions are most often made for tuples close to the decision boundary, because the prediction confidence (i.e., the probability of belonging to the predicted class) is low for those tuples. Given a classifier, Kam-Kar derives a critical region around the decision boundary and modifies the predictions for tuples in that region such that demographic parity is satisfied.

Let $f : f(\mathbb{X}, S) \rightarrow \hat{Y}$ be a classifier and $Pr(\hat{Y} \mid \mathbb{X}, S)$ be the prediction confidence. Kam-Kar defines a critical region around the decision boundary where the prediction confidence is below a threshold θ, i.e., $\max(Pr(\hat{Y} = 1 \mid \mathbb{X}, S), Pr(\hat{Y} = 0 \mid \mathbb{X}, S)) < \theta$. Here, θ is a hyper-parameter that can be tuned to find the optimal critical region for the desired level of demographic parity. Kam-Kar rejects

the predictions for tuples that belong to the critical region as those predictions are most likely to be discriminatory.

In order to enforce demographic parity, Kam-Kar modifies the predictions for the tuples in the critical region using the following method: $\hat{Y} = 1$ is assigned to all tuples belonging to the unprivileged group, while $\hat{Y} = 0$ is assigned to all tuples belonging to the privileged group.

Implementation. We collected the source code for Kam-Kar from the open source AI Fairness 360 library.[9] We set all the hyper-parameters following the instructions specified within the source code (more details are in the authors' repository).

A.3.2 Hardt. Hardt et al. [37] propose a post-processing approach that enforces equalized odds. We refer to this approach as Hardt. Given the ground truth Y and the sensitive attribute S in the training data, Hardt learns the parameters of a new mapping $g : g(\hat{Y}, S) \rightarrow \tilde{Y}$ to replace $\hat{Y}$ such that *TPR* and *TNR* are equalized across the privileged and the unprivileged groups.

In order to enforce equalized odds, the new mapping g must satisfy the following condition: $Pr(\tilde{Y} = 1 \mid S = 1, Y = y) = Pr(\tilde{Y} = 1 \mid S = 0, Y = y), \forall y \in Y$. Given any standard loss function $l : l(Y, \tilde{Y}) \rightarrow \mathbb{R}$ that quantifies the cost of incorrect predictions, Hardt solves the following linear program to obtain the optimal mapping:

$$\min_{g} \mathbb{E}[l(Y, \tilde{Y})]$$

$$\text{s.t. } Pr(\tilde{Y} = 1 \mid S = 1, Y = y) = Pr(\tilde{Y} = 1 \mid S = 0, Y = y), \forall y \in Y,$$

$$\text{and } Pr(\tilde{Y} = 1 \mid S = s, Y = y) \in [0, 1], \forall y \in Y,\ s \in S,$$

where $\mathbb{E}[l(Y, \tilde{Y})]$ is the expected loss. The solution to this linear program always provides a mapping for modifying the predictions such that equalized odds is satisfied.

Implementation. We collected the source code for Hardt from a public repository.[10]

A.3.3 Pleiss. Pleiss et al. [64] propose a post-processing approach to ensure that a calibrated classifier satisfies *equal opportunity*—equal *TPR* across the sensitive groups—or *predictive equality*—equal *FPR* across the sensitive groups—or a weighted combination thereof. We refer to this approach as Pleiss. Pleiss derives a new predictor for the group with higher *TPR* (or lower *FPR*) and replaces $\hat{Y}$ in order to enforce the fairness notion.

Pleiss begins by assuming that the optimal classifier f, learned on the training data $\mathcal{D}$, is reliable and calibrated, i.e., $Pr(Y = 1 \mid \hat{Y} = y) = y, \forall y \in Y$. Given f, Pleiss derives two cost functions, $C_0(f)$ and $C_1(f)$, for the unprivileged and the privileged groups, respectively. Depending on the target fairness notion, this cost function denotes the *TPR*, or the *FPR*, or a weighted combination thereof, for the corresponding group. f violates fairness if it favors one group, i.e., $C_0(f) \neq C_1(f)$.

To enforce the target fairness notion, Pleiss derives a new predictor for the favored group, such that it replaces a random subset of $\hat{Y}$ to decrease the *TPR* (or increase *FPR*) to make it approximately equal to the other (unfavored) group. For any tuple t in the favored group, the actual prediction $\hat{Y}_t$ is withheld with probability $\alpha \in [0, 1]$, where α depends on the difference between C_0 and C_1. Then $\hat{Y}_t$ is replaced with $\tilde{Y}_t$, such that $\tilde{Y}_t = 1$ with probability proportional to the fraction of positive tuples in the favored group. This modification technique decreases the classifier's performance for the favored group while maintaining classifier calibration, and approximately satisfies the target fairness notion.

Pleiss et al. acknowledge that their approach satisfies group-level fairness while intentionally violating individual-level fairness due to randomness in predictions.

Implementation. We collected the source code for Pleiss from the authors' public repository.[10] We use equal opportunity as the fairness notion, since minimizing the difference in terms of favorable outcomes—i.e., equal *TPR* across the sensitive groups—is more appropriate as the fairness goal in the context of our datasets. Further, a weighted combination of equal opportunity and predictive equality led to very poor performance in terms of correctness in most cases.

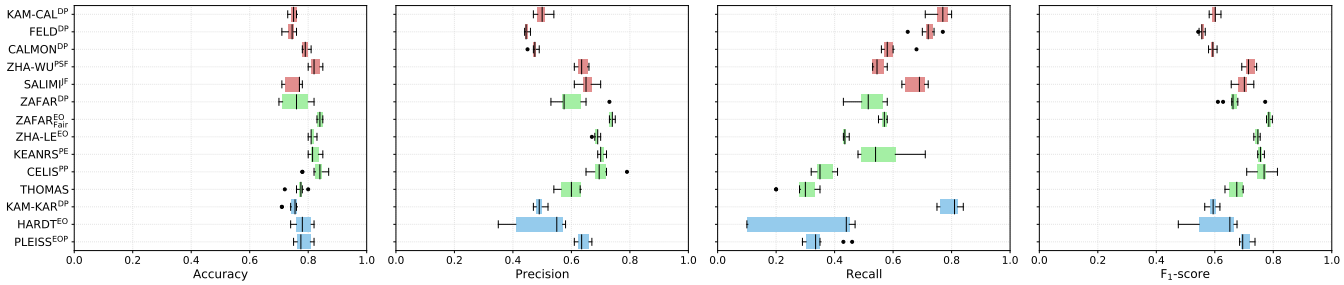
B ADDITIONAL DISCUSSION

Fairness constraints may lead to better generalization. Over the German dataset (Figure 10(c)), we observe that Celis achieves slightly better correctness than LR across all correctness metrics. Note that over this dataset, even the fairness-unaware approach LR shows low discrimination. This made it possible for Celis to make no compromise in correctness at all while enforcing fairness. As pointed out by prior work [81], fairness constraints may sometimes act as a regularizer and may lead to better generalization performance and we observe such an incident here. This indicates that enforcing fairness does not necessarily imply compromise in correctness.

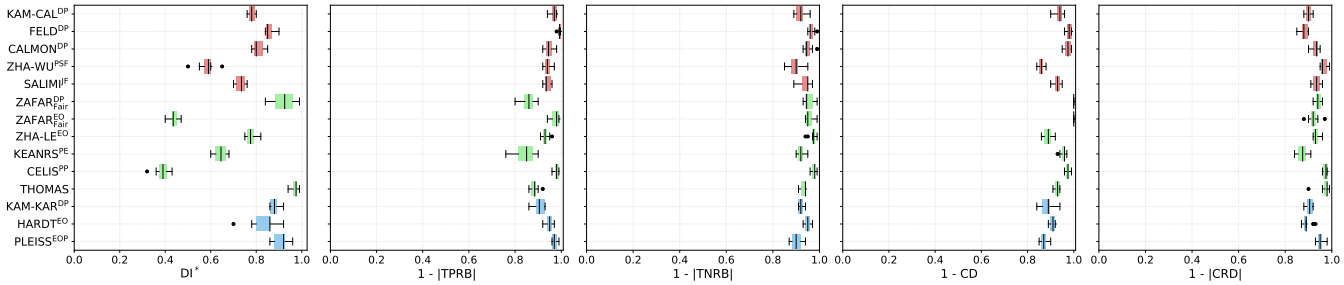
C COMPLETE RESULTS OF STABILITY EXPERIMENTS

We show the results of the stability experiments in terms of all correctness and fairness metrics over the Adult dataset in Figure 13, the COMPAS dataset in Figure 14, the German dataset in Figure 15, and the Credit dataset in Figure 16.

[9] https://github.com/Trusted-AI/AIF360/tree/master/aif360/algorithms/postprocessing
[10] https://github.com/gpleiss/equalized_odds_and_calibration

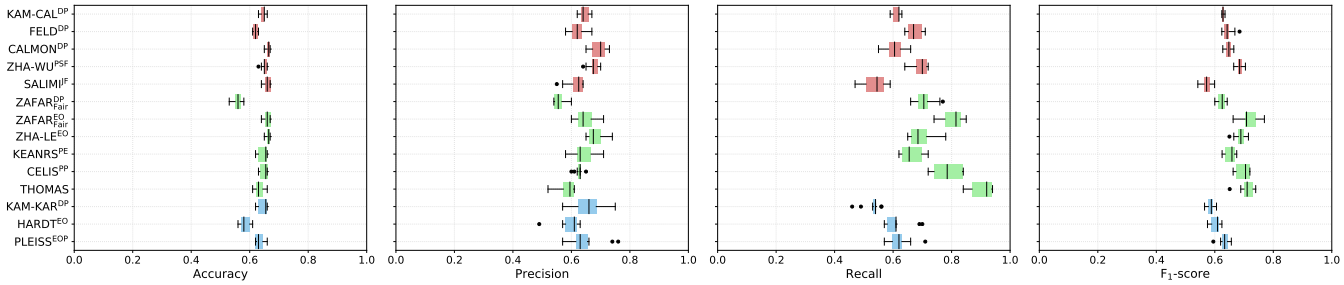


(a) Correctness

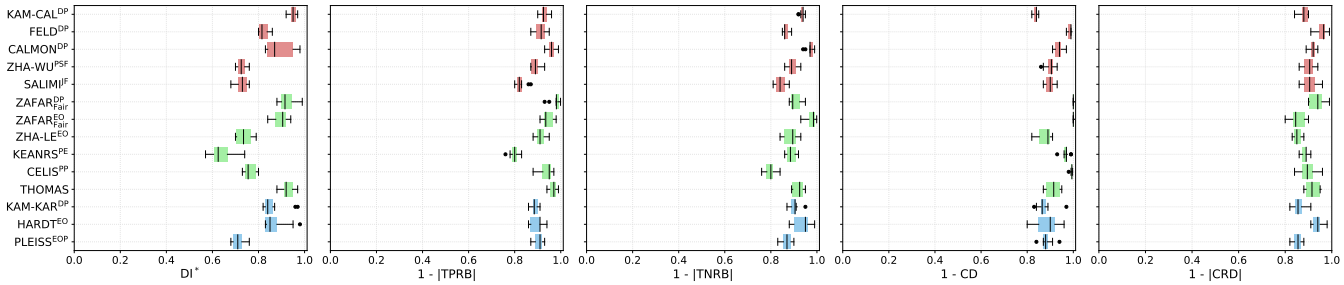


(b) Fairness

Figure 13: Variance of the fair approaches in terms of (a) correctness and (b) fairness metrics on arbitrary folds over Adult.

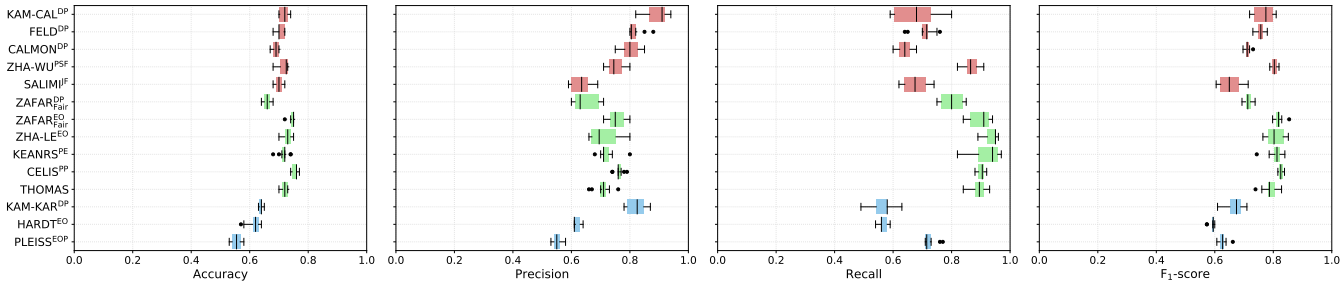


(a) Correctness

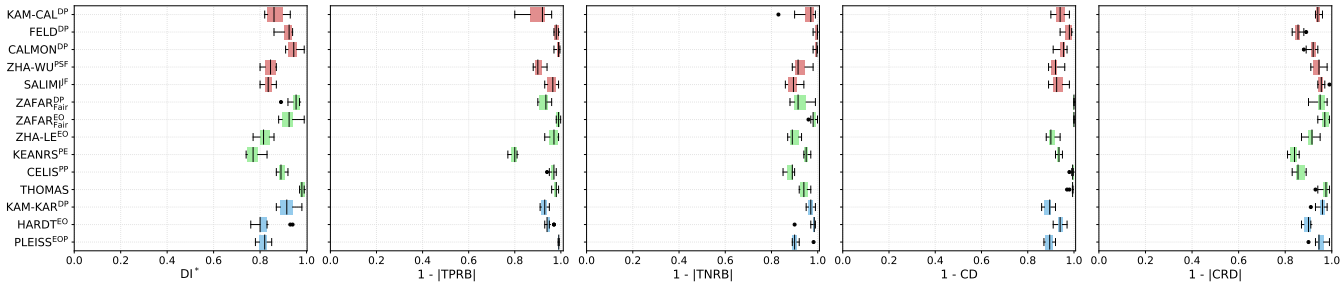


(b) Fairness

Figure 14: Variance of the fair approaches in terms of (a) correctness and (b) fairness metrics on arbitrary folds over COMPAS.

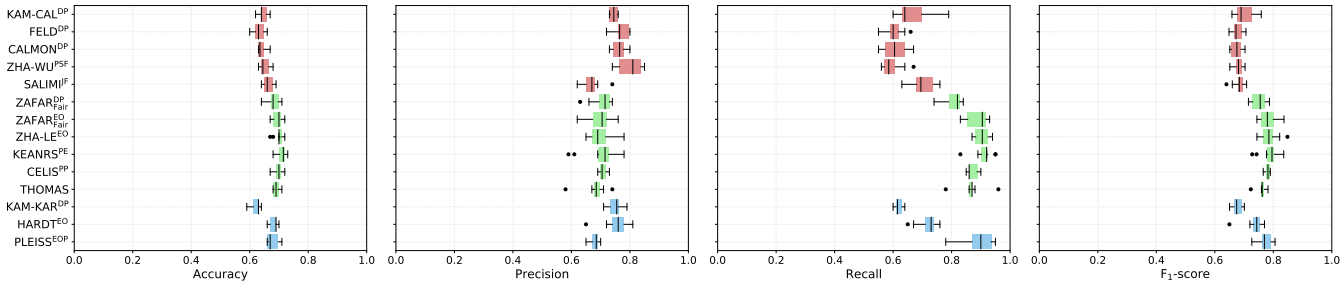


(a) Correctness

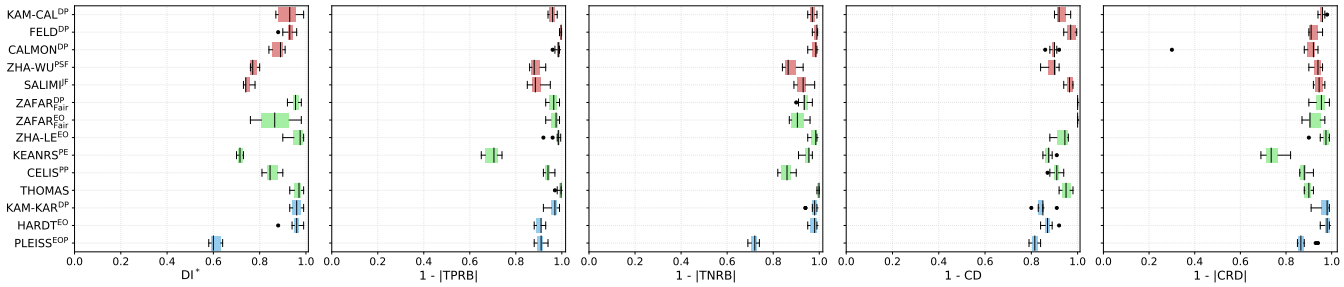


(b) Fairness

Figure 15: Variance of the fair approaches in terms of (a) correctness and (b) fairness metrics on arbitrary folds over German.



(a) Correctness



(b) Fairness

Figure 16: Variance of the fair approaches in terms of (a) correctness and (b) fairness metrics on arbitrary folds over Credit.