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

Technical Report

Maliha T Islam
University of Massachusetts Amherst
mtislam@cs.umass.edu

Anna Fariha
Microsoft
annafariha@microsoft.com

Alexandra Meliou
University of Massachusetts Amherst
ameli@cs.umass.edu

Babak Salimi
University of California, San Diego
bsalimi@ucsd.edu

ABSTRACT

Classification, a heavily studied data-driven machine learning task, drives a large number of prediction systems involving critical 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, robustness to data errors, sensitivity to underlying ML model, data efficiency, 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 [37], all the way to high-stakes decisions such as mortgage approval [19], job applicant filtering [26], and pre-trial risk assessment of criminal defendants [58]. However, automated predictions are only as good as the data that drives them. As inherent biases are common in data [6], data-driven systems commonly demonstrate unfair and discriminatory behavior [9, 58, 78, 87].

Data management research has shown growing interest in the topic of fairness over applications related to ranking, data synthesis, result diversification, and others [3–5, 33, 53, 84, 92]. 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* [27, 57, 77, 78, 98], one of the most important and heavily studied supervised ML 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 empirically 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 scalability, robustness to data errors, stability wrt to partitions of training data, and data efficiency, which are important practical considerations often overlooked in other communities, and (2) to produce a deeper understanding of tradeoffs that may exist across various approaches, creating guidelines for where data management solutions are more likely to have impact. We proceed to provide a brief 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. Classifiers typically focus on maximizing *correctness*, i.e., how well predictions match the ground truth. To that end, a trained classifier naturally prioritizes the minimization of prediction error over the majority groups within the data, and, thus, performs better for entities belonging to those groups. However, this may result in poor prediction performance over minority groups. Moreover, as all data-driven approaches, classifiers also suffer from the general phenomenon of "garbage-in, garbage-out": if the data contains inherent biases, the model will reflect or even exacerbate them. Thus, traditional learning may discriminate in two ways: (1) models make more incorrect predictions over the minority than the majority groups, and (2) they replicate training data biases. We highlight this with a real-world example.

Example 1. *Consider COMPAS, a risk-assessment system that can predict recidivism (the tendency to reoffense) in convicted criminals. It 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 [29], and achieves nearly* 70% *accuracy [24]. In 2014, a detailed analysis of COMPAS revealed some very troubling findings: black defendants are twice more likely than white defendants to be* incorrectly *predicted as high-risk, while white reoffenders are* incorrectly *predicted as low-risk almost twice as often as black reoffenders [58]. While COMPAS' overall accuracy was similar over both groups (67% for black and 69% for white), its mistakes affected the two groups disproportionately. COMPAS was further criticized for exacerbating societal bias due to training historical arrest data, despite certain populations being proven to be more policed than others [74].*

Example 1 is not an isolated incident; other cases of classifier discrimination have pointed towards racial [9], gender [78], and other forms of bias and unfairness [87]. The pervasiveness 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 [17, 27, 39, 78, 95], and a multitude of approaches and notions of fairness have emerged [67, 89]. 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. Fairness is subjective and specifying what is fair is non-trivial: definitions of fairness are often driven by application-specific and even legal considerations. Existing literature has proposed a large number of notions to capture different fairness objectives [67, 89], and new ones continue to emerge. A principled comparison of these notions is non-trivial, due to the high diversity in their mechanisms. Some fairness notions measure discrimination through *causal* association among attributes of interest (e.g., race and prediction), while others study non-causal associations. Further, some notions capture if *individuals* are treated fairly, while others quantify fair treatment of a *group* (e.g., people of certain race or gender). The demand for domain knowledge also varies: some rely on *observational* data, while others require *interventions* or *counterfactuals.* To add further complexity, multiple recent studies [23, 51, 62] prove that most fairness notions tend to be incompatible with each other and cannot be enforced simultaneously.

Fairness-enforcing stage. Existing methods in fair classification operate in one of the three possible stages. *Pre-processing* approaches attempt to repair biases in the data *before* the data is used to train a classifier [15, 27, 42, 78, 102, 103]. 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 [17, 46, 85, 93, 95, 97], and *post-processing* approaches, which alter the classifier predictions to ensure fairness [39, 44, 71]. Similar to fairness notions, the wide variety of mechanisms applied by fair approaches present a significant challenge in understanding them. Further, there is a clear lack of literature that empirically evaluate these approaches, making it difficult to compare the tradeoffs that approaches may make while enforcing fairness.

Scope of our work. We present a systematic and thorough empirical evaluation of 13 fair classification approaches and some of their variants, resulting in 18 different approaches, along axes that the data management community cares about: *correctness*, *fairness*, *scalability*, *robustness to data errors*, *sensitivity to ML model*, *data efficiency*, 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 practical concerns such as scalability, robustness to data errors, etc., that are relevant to many real-world applications but have been overlooked in fairness literature, 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. For example, our findings suggest that pre-processing approaches, while a natural fit for data-focused solutions, tend to face scalability issues with high-dimensional data. The contributions of our work lie both in the breadth of our evaluation, as well as in the unique perspective of data-management considerations, which have not been previously explored in this context. 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. Our focus on the empirical evaluation of methods in fair classification distinguishes our work from existing surveys that review the broad area but do not include experimental results and analysis [16, 63, 64, 89]. Moreover, prior work on the evaluation of fair classifiers had a narrower scope than ours. Friedler et al. [31] carry out experimental analysis similar to ours by evaluating variations of 4 fair approaches over 5 fairness metrics, while Jones et al. [41] evaluate variations of 6 fair approaches over 3 fairness metrics. However, they overlook performance aspects (e.g., runtime, scalability, data-efficiency) and robustness to data-quality issues (e.g., errors), which are critical in practice. Further, their analysis excludes post-processing approaches and individual fairness metrics. AI Fairness 360 [7] is an extensible toolkit that offers mechanisms to empirically evaluate fair approaches over different fairness metrics. However, it does not offer any insight highlighting the tradeoffs among fair approaches, and cannot compare other aspects such as efficiency, scalability, robustness to data errors, stability, etc. Lastly, a few general frameworks [32, 86] evaluate fair approaches on a specific fairness metric, but are not designed to offer insights based on comparative analysis.

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

- We provide a new and informative categorization of 34 existing fairness notions, based on the high-level aspects of association, granularity, causal hierarchy, 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* [15, 27, 42, 78, 102, 103], 5 *in-processing* [17, 46, 85, 93, 95, 97], and 3 *post-processing* approaches [39, 44, 71] 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 3 real-world datasets including Adult [52] and COMPAS [58]. 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. (Section 4.3)
- We investigate the robustness of all approaches to quality issues (e.g., errors) in training data, shedding light on their feasibility in practical settings. Our results indicate that pre- and in-processing exhibit poor generalizability and often fail to achieve their target fairness, while post-processing is more robust. (Section 4.4)

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.

- To evaluate the sensitivity of pre- and post-processing approaches to the choice of ML model, we pair each approach with 5 different ML models and compare their correctness-fairness balance. Our findings show that pre-processing approaches can produce noticeably varied results on different models, while post-processing is not sensitive to the choice of ML model. (Section 4.5)
- We summarize further results on the data efficiency (dependence on training set size) and stability (variance over different partitions of the training data) of all approaches. Our results suggest that most approaches are data-efficient and stable, and there is no significant trend. (Section 4.6)
- 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 and then 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 $t \in \mathcal{D}$. 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. We use Y_t and $\hat{Y}_t$ to denote the ground-truth and predicted class label for t, respectively. We summarize the notations in Figure 1.

2.1 Correctness

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 *TP*, *TN*, *FP*, and *FN*, which denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Further, *TPR*, *TNR*, *FPR*, and *FNR* denote the rate of true positives, true negatives, false positives, and false negatives, respectively.

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 2: List of correctness metrics used in our evaluation.

Metrics. We measure correctness through well-studied metrics in literature [55] (Figure 2). 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 when datasets have a balanced class distribution, it can be misleading for imbalanced datasets, which is found frequently in real-world scenarios. In such cases, precision, recall, and F_1-score, together, are more insightful.

2.2 Fairness

Fairness in classifier predictions typically targets sensitive attributes, such as gender, race, etc. Example 1 highlights how a classifier can discriminate despite being reasonably accurate.

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 quantifying metrics. Figure 3 presents a list of 34 fairness notions and corresponding metrics that have been studied in the literature. We primarily categorize these notions based on the association considered between the sensitive attribute and the prediction: some notions analyze the source of discrimination through *causal* relationships among the attributes, while others compute *non-causal* associations through observed statistical correlations. We highlight further distinction among the notions based on their granularity, position in the causal hierarchy, and additional requirements they impose:

Granularity. We classify fairness notions 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. Group-based notions can further be categorized as *demography-aware*, which consider the distribution of outcomes among groups to measure fairness, and *error-aware*, which compare the error rates for each group.

Causal hierarchy. A key feature of the fairness notions is their position in the causal hierarchy that is determined by the extent of domain knowledge they require. We highlight this distinction using

	Fairness notion	Metric	Granularity			Causal hierarchy			Additional requirements			
			group		individual	observation	intervention	counterfactual	prediction probability	causality model	resolving attribute	similarity metric
			demography-aware	error-aware								
non-causal	conditional statistical parity [23]	conditional statistical parity	✓			✓						
	demographic parity[†] [25]	disparate impact [95], CV score [13]	✓			✓						
	intersectional fairness [30]	differential fairness	✓			✓						
	conditional accuracy equality [9]	false discovery/omission rate parity		✓		✓						
	predictive parity [22]	false discovery rate parity		✓		✓						
	overall accuracy equality [9]	balanced classification rate [31]		✓		✓						
	treatment equality [9]	ratio of false negative and false positive		✓		✓						
	equalized odds [39]	true positive/negative rate balance		✓		✓						
	equal opportunity[‡] [39]	true negative rate balance		✓		✓						
	resilience to random bias [28]	resilience to random bias		✓		✓						
	preference-based fairness [94]	group benefit		✓		✓						
	calibration [22]	calibration		✓		✓			✓			
	calibration within groups [51]	well calibration		✓		✓			✓			
	positive class balance [51]	fairness to positive class		✓		✓			✓			
	negative class balance [51]	fairness to negative class		✓		✓			✓			
	individual discrimination[††] [32]	individual discrimination			✓	✓						
	metric multifairness [50]	metric multifairness			✓	✓						✓
	fairness through awareness [25]	fairness through awareness			✓	✓						✓
	fairness through unawareness [54]	Kusner et al. [54]			✓	✓						
causal	proxy fairness [48]	proxy fairness	✓				✓			✓		
	total causal effect [70]	total effect	✓				✓			✓		
	direct causal effect [70]	natural direct effect	✓				✓			✓		
	indirect causal effect [70]	natural indirect effect	✓				✓			✓		
	path-specific fairness [103]	path specific effect	✓				✓			✓		
	unresolved discrimination [48]	causal risk difference [73]	✓				✓				✓	
	interventional/justifiable fairness [78]	ratio of observable discrimination	✓				✓				✓	
	fair on average causal effect [47]	fair on average causal effect	✓				✓			✓		
	non-discrimination criterion [102]	non-discrimination criterion	✓				✓			✓		
	equality of effort [40]	equality of effort	✓				✓			✓		
	counterfactual effects [100]	counterfactual direct/indirect effect	✓					✓		✓		
	counterfactual error rates [99]	counterfactual error rates		✓				✓		✓		
	counterfactual fairness [54]	counterfactual effect [91]			✓			✓		✓		
	path-specific counterfactuals [91]	counterfactual effect			✓			✓		✓		
	individual direct discrimination [101]	individual direct discrimination			✓		✓			✓		

Figure 3: We categorize fairness notions and metrics in the literature according to the type of association considered between attributes (causal or non-causal) and list other properties: granularity (group or individual), and position in the causal hierarchy based on required domain knowledge (observation, intervention, or counterfactual). Here, we use intervention in the context of causal inference that requires interventions to adhere to the causal model of the data. We further partition group-level notions based on their strategy to measure discrimination (demography- or error-aware). All notions require knowledge of the sensitive attributes and the classifier predictions. Some definitions place additional requirements, shown in the right-most four columns. For our evaluation, we choose five metrics (Figure 4) that cover the highlighted definitions. ([†]also known as statistical parity; [‡]also known as predictive equality; [††]also known as causal discrimination).

Pearl's ladder of causation [70], a hierarchy of three levels of increasing complexity: (1) observation (2) intervention, and (3) counterfactual. Notions at the *observation* level can be computed entirely from observational data. Notions at the *intervention* level use both observational data and the underling causal structure, i.e., an abstract model that shows whether any causal relationship exists between attributes. Lastly, notions at the *counterfactual* level demand observational data, and full specification of the underlying causal model denoting the exact functional relationships between attributes.

Additional requirements. All notions require information on the sensitive attribute and the classifier predictions. Some notions impose additional requirements, such as causality models or causal structure [70], resolving attributes that mediate the relationship between the sensitive attribute and the outcome in non-discriminatory ways [73], similarity metric between individuals [25], etc.

2.2.2 *Fairness metrics.*

While Figure 3 highlights a wide range of proposed fairness notions, Prior works [31, 62] 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 4) that are most prevalent in the literature and capture commonly occurring discriminations in binary classification [22]. We briefly review these metrics and refer to the Appendix for a detailed discussion.

Non-causal metrics depend entirely on empirical data and look for statistical relationships between the sensitive attribute and the prediction. We experiment with the following non-causal metrics:

Disparate Impact (DI) compares the distribution of predictions among sensitive groups and captures if they are independent of the sensitive attribute. Specifically, *DI* computes the ratio of empirical probabilities of receiving positive predictions between the unprivileged and the privileged groups (Figure 4, row 1). *DI* is also commonly known by its corresponding notion, demographic parity [25].

True Positive Rate Balance (TPRB) and True Negative Rate Balance (TNRB) measure discrimination as the difference in *TPR* and *TNR*, respectively, between the privileged and the unprivileged groups (Figure 4, rows 2–3). These metrics are also known as equalized odds [39], the notion they jointly measure.

Individual Discrimination (ID) [32] checks whether assigning different values to the sensitive attribute changes the prediction for an individual. Specifically, *ID* is computed as the fraction of tuples for which changing the sensitive attribute causes a change in the prediction for otherwise identical data points (Figure 4, row 4).

Causal metrics determine discrimination by considering the causal relationship between the sensitive attribute and the prediction, as opposed to their statistical dependencies. As non-causal metrics cannot reason about whether a sensitive attribute is the true cause of discrimination, causal metrics address this limitation through additional domain knowledge. We experiment with

Total Effect (TE) [70], a causal metric that measures discrimination as the causal influence of the sensitive attribute on prediction. It measures the effect of interventions to the sensitive attribute on the prediction, to determine the extent of causal influence (Figure 4, row 5). *TE* is often decomposed into indirect (causal influence mediated by other attributes) and direct (influence that is not mediated) effects, or path-specific effects (influence through particular causal pathways) that are needed in many real-world situations [1, 99].

Discussion on metric choices. We select metrics to cover a variety of categories in our classification, including causal and non-causal associations, group- and individual-level fairness, and observational and interventional techniques (highlighted rows in Figure 3). Other causal notions can also address the limitations of non-causal metrics, but they often require additional information (e.g., structural equations for counterfactuals) to be computed from observational data. We choose metrics that are feasible within the scope of our experiments, and exclude ones that make strong and impractical assumptions about the problem setting [70]. For similar reasons, we do not include individual-level metrics that depend on counterfactuals or similarity measures between individuals.

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. For our evaluation, we select approaches that span all three stages and target a representative variety of fairness notions, including causal and non-causal associations, observation- and intervention-level techniques. Figure 5 overviews our chosen approaches. We proceed to provide a high-level description of the approaches in each category, underscoring their similarities and differences. (More details are in the Appendix).

3.1 Pre-processing

Pre-processing approaches are motivated from the fact that ML techniques are data-driven and the predictions of a classifier reflect trends and biases in 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 made by a learned classifier satisfy the target fairness notion. The main advantage of pre-processing is that it is model-agnostic, allowing 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 do not always come with provable guarantees of fairness.

In our evaluation, we include three pre-processing approaches that enforce non-causal fairness notions and two approaches that target causal notions. We briefly discuss these approaches here.

Kam-Cal [42] is a pre-processing approach that enforces demographic parity, a notion that ensures model prediction $\hat{Y}$ is independent of the sensitive attribute S. Assuming that $\hat{Y}$ reasonably approximates the ground truth Y, Kam-Cal argues that $\hat{Y}$ is likely to be independent of S when the classifier is deployed, if there is no dependency between Y and S in the training data. To this end, Kam-Cal resamples the training data $\mathcal{D}$ with a weighted sampling technique to ensure that S and Y are statistically independent.

Feld [27] is another approach that enforces demographic parity. It argues that demographic parity can be ensured if the marginal distribution of each attribute $X \in \mathbb{X}$ is similar across the sensitive groups in training data $\mathcal{D}$. Intuitively, if a model learns from such data, it is likely to predict based on attributes that are independent of S, which in turn satisfies demographic parity. To that end, Feld modifies the values for each attribute X until the marginal distributions are similar for the privileged and unprivileged group. Unlike Kam-Cal that only resamples the tuples, Feld modifies the training data. Further, Kam-Cal enforces demographic parity through independence between S and Y, while Feld reformulates it as an independence condition between $\mathbb{X}$ and S.

Calmon [15] is one more approach targeting demographic parity. The goal of this approach is to reduce the dependency between S and Y by minimally perturbing the attribute values of $\mathbb{X}$ and Y and without significantly distorting the underlying data distribution. It utilizes the joint distribution associated with $\mathcal{D}$ and a set of pre-defined distortion functions to define the corresponding optimization problem for minimal repair. Calmon uses convex optimization techniques to solve this optimization problem and minimally modifies $\mathbb{X}$ and Y to achieve the target fairness goal. Among Kam-Cal, Feld, and Calmon, it is the only approach that modifies both training and test data.

Zha-Wu [102, 103] proposes two methods that target causal notions: path-specific fairness (Zha-Wu$^{\text{PSF}}$) and direct causal effect

Metric	Definition	Fairness notion	Range	Interpretation
Disparate Impact (*DI*) [27]	$\frac{Pr(\hat{Y}=1 \mid S=0)}{Pr(\hat{Y}=1 \mid S=1)}$	demographic parity	[0, ∞)	$DI = 1 \longrightarrow$ completely fair $DI = 0 \longrightarrow$ completely unfair $DI = \infty \longrightarrow$ completely unfair
True Positive Rate Balance (*TPRB*) [39]	$Pr(\hat{Y}=1 \mid Y=1, S=1) - Pr(\hat{Y}=1 \mid Y=1, S=0)$	equalized odds	[-1, 1]	$\lvert TPRB \rvert = 0 \longrightarrow$ completely fair $\lvert TPRB \rvert = 1 \longrightarrow$ completely unfair
True Negative Rate Balance (*TNRB*) [39]	$Pr(\hat{Y}=0 \mid Y=0, S=1) - Pr(\hat{Y}=0 \mid Y=0, S=0)$	equalized odds	[-1, 1]	$\lvert TNRB \rvert = 0 \longrightarrow$ completely fair $\lvert TNRB \rvert = 1 \longrightarrow$ completely unfair
Individual Discrimination (*ID*) [32]	$\frac{\lvert Q \rvert}{\lvert \mathcal{D} \rvert}$, given $Q = \{a \in \mathcal{D} \mid \exists b : \mathbb{X}_a{=}\mathbb{X}_b \land S_a{\neq}S_b \ \land \hat{Y}_a{\neq}\hat{Y}_b\}$	individual discrimination	[0, 1]	$ID = 0 \longrightarrow$ completely fair $ID = 1 \longrightarrow$ completely unfair
Total Effect (*TE*) [70]	$Pr(\hat{Y}_{S=1} = 1) - Pr(\hat{Y}_{S=0} = 1)$	total causal effect	[-1, 1]	$\lvert TE \rvert = 0 \longrightarrow$ completely fair $\lvert TE \rvert = 1 \longrightarrow$ completely unfair

Figure 4: List of fairness metrics we use to evaluate fair classification approaches. These metrics effectively contrast between causal and non-causal associations; and cover group- and individual-level discrimination, observation- and intervention-level techniques.

(Zha-Wu$^{\text{DCE}}$). Zha-Wu$^{\text{PSF}}$ enforces path-specific fairness by modifying Y such that all causal influences of S over Y are removed. It learns a causal graph over $\mathcal{D}$ to discover (direct and indirect) causal associations between Y and S, and translates the minimal repair of Y to a quadratic programming problem. On the other hand, Zha-Wu$^{\text{DCE}}$ aims to minimize the direct causal effect of S on Y. It determines a set of parents (Q) of Y that blocks all indirect causal paths from S to Y and uses Q to compute the causal effect on the direct path. Then, it modifies the distribution of Y such that the direct causal effect is below a user-defined threshold. Zha-Wu is different from all the aforementioned approaches as it enforces causal notions using additional domain knowledge.

Salimi [78] enforces justifiable fairness, a causal notion that prohibits causal dependency between S and $\hat{Y}$, except through a set of admissible attributes $\mathbf{A} \in \mathbb{X}$. $\mathbf{A}$ is pre-defined such that the effect of S on $\hat{Y}$ through $\mathbf{A}$ is deemed non-discriminatory. All other attributes are considered discriminatory and constitute the inadmissible set ($\mathbf{I}$). Similar to other approaches, Salimi assumes that $\hat{Y}$ is likely to be fair if a classifier is trained on data $\mathcal{D}$ where Y satisfies the target fairness notion. It enforces justifiable fairness as a conditional independence and minimally modifies the underlying data distribution such that Y is conditionally independent of $\mathbf{I}$ given $\mathbf{A}$. Salimi solves the optimization problem corresponding to minimal repair using weighted maximum satisfiability (Salimi$^{\text{JF}}_{\text{MaxSAT}}$) and matrix factorization (Salimi$^{\text{JF}}_{\text{MatFac}}$). Unlike Zha-Wu, Salimi does not require the entire causal graph and repairs $\mathcal{D}$ only by inserting or deleting tuples.

3.2 In-processing

In-processing approaches are most favored by the machine learning community [17, 46, 95, 97] 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 choose five different in-processing approaches and their variants, to best highlight the variety of techniques that exist in literature. We provide a concise summary of these approaches next.

Zafar [93, 95] proposes two methods to enforce demographic parity (Zafar$^{\text{DP}}$) and equalized odds (Zafar$^{\text{EO}}_{\text{Fair}}$). Both of these approaches utilize tuples' distance from the decision boundary as a proxy of $\hat{Y}$ to model fairness violations, translate their corresponding fairness notion to a convex function of the classifier parameters. Zafar$^{\text{DP}}$ solves the resulting constrained optimization problem to compute the optimal classifier parameters that either maximizes prediction accuracy under fairness constraints (Zafar$^{\text{DP}}_{\text{Acc}}$), or minimizes fairness violation under constraints on accuracy compromise (Zafar$^{\text{DP}}_{\text{Fair}}$). On the other hand, Zafar$^{\text{EO}}_{\text{Fair}}$ only computes parameters that maximize prediction accuracy under fairness constraint through a disciplined convex-concave program [83].

Zha-Le [97] enforces the notion of equalized odds. It leverages *adversarial learning*, a technique where a classifier and an adversary with mutually competing goals are trained together. Given $\mathcal{D}$, the goal of the classifier is to maximize the accuracy of $\hat{Y}$, while the adversary attempts to correctly predict S using $\hat{Y}$ and Y. Zha-Le utilizes gradient descent techniques [11] to compute the classifier's optimal parameters such that $\hat{Y}$ does not contain any information about S that the adversary can exploit.

Kearns [46] is an in-processing approach that either enforces demographic parity, or predictive equality (i.e., equal *FPR* for the sensitive groups). Kearns aims to approximately enforce the target fairness notion within a set of subgroups defined using one or more (user-specified) sensitive 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 entire population.

Celis [17] accommodates a wide range of notions: predictive parity, demographic parity, equalized odds, and conditional accuracy equality. It reduces all fairness notions to linear forms and solves the corresponding convex optimization problem using Lagrange multipliers [38] to minimize prediction error under fairness constraints. Unlike prior approaches that only enforce specific fairness notions, Celis is designed to support a wide variety of fairness notion within a single framework.

Stage	Approach	Fairness notion(s)	Key mechanism	Evaluated version(s)
pre	Kam-Cal [42]	demographic parity	Apply weighted resampling over tuples in $\mathcal{D}$ to remove dependency between S and Y.	• Kam-Cal$^{\text{DP}}$
	Feld [27]	demographic parity	Repair each $X \in \mathbb{X}$ independently s.t. X's marginal distribution is indistinguishable across sensitive groups. Training and test data are both modified.	• Feld$^{\text{DP}}$
	Calmon [15]	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. Training and test data are both modified.	• Calmon$^{\text{DP}}$
	Zha-Wu [102, 103]	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$^{\text{PSF}}$
		direct causal effect	Given a causal graph, identify the set of parents (Q) of Y that blocks all indirect paths from S to Y. Use Q to compute the causal effect of S on Y through the direct path and modify Y s.t. this effect is within allowable threshold.	• Zha-Wu$^{\text{DCE}}$
	Salimi [78]	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$^{\text{JF}}_{\text{MaxSAT}}$ (Weighted maximum satisfiability) • Salimi$^{\text{JF}}_{\text{MatFac}}$ (Matrix factorization)
in	Zafar [93, 95]	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$^{\text{DP}}_{\text{Fair}}$ (Maximize accuracy under constraint on demographic parity) • Zafar$^{\text{DP}}_{\text{Acc}}$ (Maximize demographic parity under constraint on accuracy) • Zafar$^{\text{EO}}_{\text{Fair}}$ (Same as Zafar$^{\text{DP}}_{\text{Fair}}$, but use misclassified tuples only)
	Zha-Le [97]	equalized odds	Utilize adversarial learning to train 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$^{\text{EO}}$
	Kearns [46]	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$^{\text{PE}}$ (For subgroups $\{\mathcal{D}_1, \mathcal{D}_2, \ldots\}$ where each $\mathcal{D}_i \subset \mathcal{D}$, ensure that $\forall \mathcal{D}_i$, $FPR(\mathcal{D}_i) \approx FPR(\mathcal{D})$)
	Celis [17]	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$^{\text{PP}}$ (Enforce $Pr(Y{=}0 \mid \hat{Y}{=}1, S{=}0) \approx Pr(Y{=}0 \mid \hat{Y}{=}1, S{=}1)$)
	Thomas [85]	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$^{\text{DP}}$ (Enforce demographic parity) • Thomas$^{\text{EO}}$ (Enforce equalized odds)
post	Kam-Kar [44]	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$^{\text{DP}}$
	Hardt [39]	equalized odds	Derive new predictor based on $\hat{Y}$ and S s.t. *TPR* and *TNR* are similar across sensitive groups.	• Hardt$^{\text{EO}}$
	Pleiss [71]	equal opportunity predictive equality	Modify $\hat{Y}$ for random tuples to equalize *TPR* (or *FPR*) across sensitive groups.	• Pleiss$^{\text{EOP}}$ (Equalize *TPR*)

Figure 5: 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.

Thomas [85] is another approach that provides a general framework to accommodate a large number of notions. It supports demographic parity, equalized odds, equal opportunity, and predictive equality. Given $\mathcal{D}$ 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.

3.3 Post-processing

Post-processing approaches are model-agnostic and enforce fairness by manipulating the predictions made by an already-trained classifier. 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 briefly describe the techniques behind the three post-processing approaches we evaluate.

Kam-Kar [44] targets demographic parity 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 randomly modifies $\hat{Y}$ for tuples in that region until the probability of positive outcome is similar across sensitive groups, i.e., demographic parity is achieved.

Hardt [39] enforces equalized odds through modifying the predictions $\hat{Y}$. Given access to Y and S from the training data $\mathcal{D}$, 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 sensitive groups. The new mapping is learned by solving a linear program.

Pleiss [71] enforces equal opportunity (equal *TPR* across the sensitive groups) or predictive equality (equal *FPR* across the sensitive groups), while maintaining the consistency (i.e., calibration) between the classifier's prediction probability for a class with the expected frequency of that class. To that end, Pleiss modifies $\hat{Y}$ for a random subset of tuples within the group with higher *TPR* (or lower *FPR*) until *TPR* (or *FPR*) is equalized.

Other approaches. We evaluate and discuss a few more approaches (not listed in Figure 5) in the Appendix. While other fair classification approaches exist, some are incorporated in the ones we evaluate [13, 14, 43], while others are empirically inferior [45], offer weaker guarantees [2, 72], do not offer a practical solution [69, 90], or do not apply to the classification setting [36, 56, 60, 79]. Some make strong assumptions about the problem setting [21, 49, 54, 65, 66, 76, 99, 100], or require additional information [25, 57, 75, 96], 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 5. The objectives of our performance evaluation are: (1) to contrast the effectiveness of all approaches in enforcing fairness and observe correctness-fairness tradeoffs, i.e., the compromise in correctness to achieve fairness (Section 4.2), (2) to contrast their efficiency and scalability with varying dataset size and dimensionality (Section 4.3), (3) to compare robustness against errors in training data (Section 4.4), (4) to compare the sensitivity of pre- and post-processing approaches to the choice of ML models (Section 4.5), and (5) to contrast stability (lack of variability) over different partitions of training data and to contrast data efficiency (dependence on dataset size) of all approaches (Section 4.6). 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 all approaches behave unpredictably in the presence of corrupt data; however, post-processing is generally more robust than pre-processing and in-processing.

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 of 13 fair classification approaches (Figure 5). Pre- and post-processing approaches require 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 approaches. Moreover, to contrast all the fair approaches against a fairness-unaware approach, we trained an unconstrained logistic regression classifier (LR) over each dataset. Hyper-parameter settings of all approaches are detailed in the Appendix.

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 code from the authors' public repositories, some by contacting the authors, and the rest from the open source library AI Fairness 360 [7] (additional details are in the Appendix). All 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 used the open source library DoWhy [82] to compute causal quantities. We implemented the evaluation scripts in Python 3.6.[1]

Metrics. We evaluated all approaches using four correctness metrics (Figure 2) and five fairness metrics (Figure 4). 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* ($DI \rightarrow 0$ and $DI \rightarrow \infty$) is mapped to low values for DI^*. Further, we report $1 - |TPRB|$, $1 - |TNRB|$, $1 - ID$, $1 - |TE|$; this way, high discrimination with respect to, say, *TPRB*, maps to low fairness value in $1 - |TPRB|$. Moreover, *ID* requires two parameters: a confidence fraction and an error-bound. We choose a confidence of 99% and an error-bound of 1%, which implies that discrimination computed using *ID* is within 1% error margin of the actual discrimination with 99% confidence.

Datasets. Our evaluation includes 3 real-world datasets, summarized in Figure 6. 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 [31, 41, 64].

Adult [52] is extracted from the 1994 US census and contains information about individuals over demographic and occupational attributes such as race, sex, education level, occupation, etc. Adult reflects historical gender-based income inequality: 11% of the females report high income ($Y = 1$), 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 [58] contains background information—such as age, sex, prior convictions, etc.—of defendants arrested in 2013–2014 and their subsequent assessment scores by the COMPAS recidivism tool [24]. The data contains racial bias: 51% African-Americans re-offend within two years ($Y = 0$), compared to 39% in other races. We select `race` as the sensitive attribute with `African-American` as the unprivileged and all other races as the privileged group.

German [34] contains records of individuals applying for credit or loan to a bank, with attributes age, sex, credit history, savings, etc. 70% of the entire population are of low credit risk ($Y = 1$), with this percentage being 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.

[1] https://github.com/maliha93/Fairness-Analysis-Code

Dataset	Size (MB)	$\lvert\mathcal{D}\rvert$	$\lvert\mathbb{X}\rvert$	S	Sensitive groups		Target task
					Unprivileged	Privileged	
Adult	3.70	45,222	9	Sex	Female	Male	Income ≥ \$50K
COMPAS	0.18	7,214	3	Race	African-American	Others	Risk of recidivism
German	0.04	1,000	9	Sex	Female	Male	Credit risk

Figure 6: 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.

Train-validation-test setting. The train-test split for each dataset was 70%-30% (using random selection) and we validated each classifier using 5-fold cross validation.

4.2 Correctness and Fairness

Figure 7 presents our correctness and fairness results over all approaches and metrics across the 3 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 7(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$^{\text{DP}}$ and Calmon$^{\text{DP}}$) demonstrate a much larger accuracy drop than the approaches that target equalized odds (such as Zafar$^{\text{EO}}_{\text{Fair}}$, Zha-Le$^{\text{EO}}$, and Kearns$^{\text{PE}}$). Zafar$^{\text{DP}}_{\text{Acc}}$ 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$^{\text{EO}}$ and Pleiss$^{\text{EOP}}$, 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 two 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.

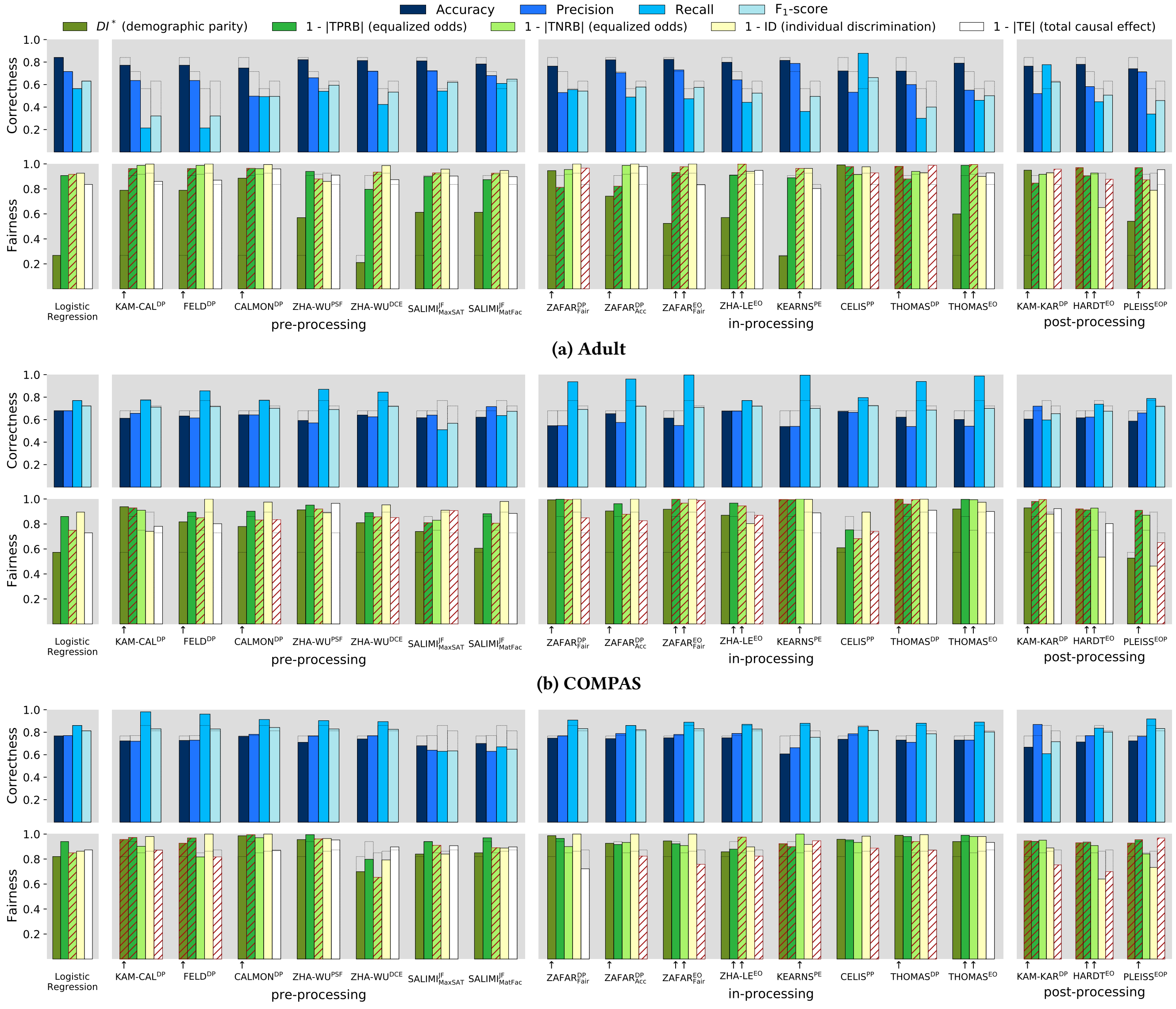
There is no single winner. All approaches succeed in improving fairness wrt the metric (and notion) they target. However, they cannot guarantee fairness wrt other notions: their performance wrt those notions is generally unpredictable. This is in line with the impossibility theorem, which states that enforcing multiple notions of fairness is impossible in the general case [22]. 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$^{\text{EO}}$ comes close in the German dataset, but this dataset contains low gender bias as even LR achieves reasonable fairness scores on all metrics, especially compared to Adult and COMPAS. Further, 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.

Causal fairness metrics explain some of the apparent discrimination. We noted a significant difference in the proportions of *TE* that transmit through the direct and indirect paths on Adult (detailed in the Appendix), which signifies that the causal influence of gender on outcome mostly goes through indirect paths. Specifically, attributes such as education, occupation, working hours/week, etc. mediate this causal influence and partially explain why there is an income gap between genders. We observe that all causal approaches, Zha-Wu$^{\text{PSF}}$, Zha-Wu$^{\text{DCE}}$, and Salimi$^{\text{JF}}$, consistently improve the fairness scores in *TE* over all datasets. In contrast, non-causal approaches behave unpredictably and often decrease the scores in *TE*, particularly the component that controls direct (and discriminatory) causal influence of the sensitive attribute.

Key takeaway: Reasoning about the causal structure is important, as it provides useful clues in understanding and explaining discrimination. Non-causal approaches establish statistical balance at the cost of exacerbating causal biases. We are not arguing that *TE* alone can resolve biases; arguably, the fact that women earn less due to their education and occupation may in itself be a bias we want to eliminate. More fine-grained causal notions are needed to capture the nuances of fairness in a particular setting.

Post-processing approaches tend to violate individual level fairness. We note that the fairness scores in *ID* are generally lower for post-processing approaches than pre- and in-processing. This is because post-processing operates on less information than pre- and in-processing and does not assume knowledge of the attributes in the training data. Thus, it does not take similarity of individuals into account and tends to produce different outcomes based on the sensitive attribute. However, some pre- and in-processing approaches—such as Feld$^{\text{DP}}$, Zafar$^{\text{DP}}$, and Zafar$^{\text{EO}}_{\text{Fair}}$—trivially satisfy *ID* by discarding the sensitive attribute while training, even though they do not target individual fairness. This indicates that *ID* is too rigid to fully capture individual discrimination as it only compares identical (except for the sensitive attribute) rather than similar individuals.



(a) Adult

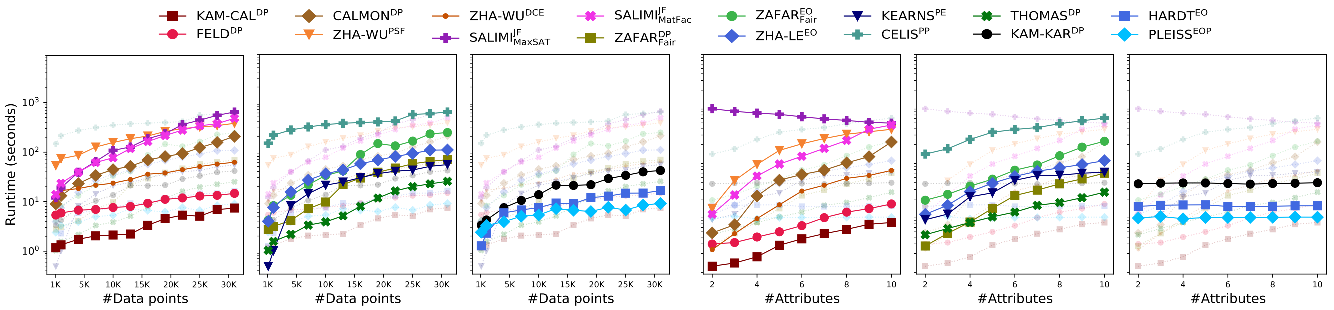
(b) COMPAS

(c) German

Figure 7: Correctness and fairness scores of the 18 fair classification approaches over (a) Adult, (b) COMPAS, and (c) German 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.

Key takeaway: Post-processing approaches can significantly violate individual level fairness. 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. However, *ID* is too rigid in practice and higher fairness scores in *ID* among pre- and in-processing approaches do not necessarily translate to higher individual fairness.

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 of 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 [90], post-processing approaches typically achieve inferior correctness-fairness balance compared to other approaches. In all the datasets, post-processing



(a) pre-processing **(b) in-processing** **(c) post-processing** **(d) pre-processing** **(e) in-processing** **(f) post-processing**

Figure 8: Results of efficiency and scalability experiments on the fair approaches. (a) – (c) show runtime overhead with varying data size and (d) – (f) show runtime overhead with varying number of attributes in Adult dataset. Note that the y-axis is in log scale.

achieves on average 2-5% lower accuracy compared to pre- and in-processing that target the same fairness metrics. There is no significant difference in performance among the pre- and in-processing approaches. However, we note that the correctness-fairness balance of pre-processing approaches varies depending on the downstream ML model (Section 4.5), and, thus, we cannot conclude if pre-processing is always comparable in performance with in-processing.

> *Key takeaway:* Pre- and in-processing achieve better correctness and fairness compared to post-processing. The performance of pre- and in-processing approaches is not always comparable as the former varies depending on the choice of ML model.

4.3 Efficiency and Scalability

In this section, we study the runtime behavior of all approaches, to investigate their efficiency gap and highlight the need for scalability considerations. While several approaches can benefit from optimizations, such as the use of GPUs, producing these optimizations is beyond the scope of our work. 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 preprocessing 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 approaches as the number of data points increases. We used the Adult dataset, as it contains the highest number of data points, and executed new instances of each approach with different numbers of data points (from 0.1K to 31K) sampled from the dataset. Our second experiment explores the runtime behavior of the approaches as the number of attributes increases. We used the Adult dataset here as well, as it contains the highest number of attributes. We executed new instances of each approach with different number of attributes (from 2 to 10). We present the results in Figure 8.

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 Thomas^{DP} 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 sharper runtime penalties. An important observation from Figure 8(a) is that causal mechanisms—such as Zha-Wu^{PSF}, Zha-Wu^{DCE}, and Salimi^{JF}—incur 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 8(d)).

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

Pre-processing scales better with increasing data sizes than with increasing number of attributes. We note a clear separation between the inherently more complex pre-processing methods (Zha-Wu^{PSF}, Zha-Wu^{DCE}, 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 in terms of efficiency, and generally better than most in-processing approaches. Generally, pre-processing demonstrates more robust scaling behavior wrt 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 8(d)). Causality-based approaches display similar challenges. The behavior of Salimi$^{JF}_{MaxSAT}$ is of note: in contrast with other techniques, its performance improves as the number of attributes grows. This is because the number of constraints in Salimi$^{JF}_{MaxSAT}$ increases rapidly with fewer attributes, resulting in higher runtimes in those settings.

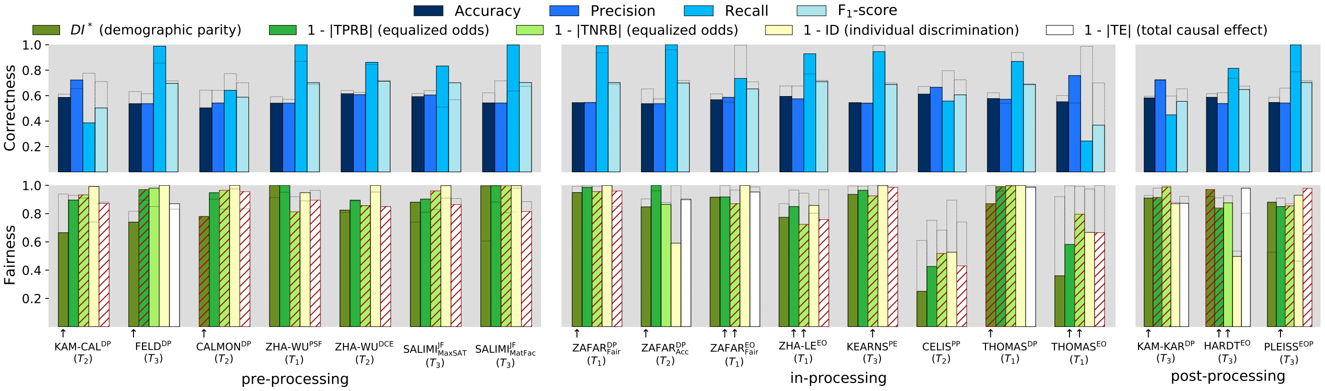


Figure 9: We examine the robustness of fair approaches to data errors. Higher scores for correctness (fairness) metrics correspond to more correct (fair) outcomes. The bars highlighted in red denote the reverse direction of the remaining discrimination—favoring the unprivileged group more than the privileged group. The arrows (↑) denote the fairness metric(s) each approach is optimized for. The bar plots for each approach on the error-free dataset are overlaid for aiding visual comparison.

In-processing approaches are more affected by the data size than by the number of attributes, but the difference is less distinct than pre-processing. In-processing techniques show a slightly sharper rise in runtime when the data size increases compared to pre-processing approaches (Figure 8(b)) and 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 appear to scale better with the number of attributes than with the data size, but this distinction is less clear than pre-processing.

4.4 Robustness to Data Errors

Fair ML approaches typically assume (explicitly or implicitly) that training and testing data are drawn from the same target distribution; thus, they can only address discrimination that is reflected in the data generative process. However, training data is susceptible to data quality issues such as selection bias, misclassification, technical errors, etc., which are introduced during data collection and preparation, and distort the underlying distribution in a way that data no longer represents the target population [80, 81]. Furthermore, data quality issues are highly correlated with sensitive attributes in many domains like healthcare and immigration [18, 68]. For example, African-American patients are more likely to be seen in clinics where documentation is less accurate or systematically different than other higher-end healthcare services [35].

In this section, we investigate the robustness of fair ML approaches to data quality issues. For this experiment, we injected COMPAS with various combinations of common data errors; we present our findings on three training datasets that contain the following errors: (T_1) swapped values between `Prior_convictions` and `Age`; (T_2) scaled values of `Prior_convictions` and noisy values of `Age`; (T_3) missing values of `Race` and `Risk_of_recidivism` that are imputed using standard Scikit-learn imputers. All errors were randomly and disproportionately introduced, affecting 50% of African-Americans and 10% of other races. The main purpose of our experiments is to highlight situations where classifiers may perform unexpectedly, not to exhaustively evaluate over all possible scenarios. We present the results in Figure 9; for each approach, we only report our findings on the set that most affected the correctness-fairness balance and refer to the Appendix for full results.

Post-processing approaches are more robust against data errors than pre- and in-processing. Post-processing is designed to manipulate the predictions of a learned classifier and does not access the data attributes. Hence, our experiments with T_1 and T_2 did not significantly affect the fairness scores of post-processing approaches. We find that post-processing approaches are most affected when trained on T_3, as they rely on the sensitive attribute and labels in the training data. We notice 2-5% drop in accuracy and F1-score, and 5-10% decrease in the fairness metrics the approaches optimize.

Pre- and in-processing methods are affected by all types of data errors. In most cases, we see a sharp decline in accuracy ranging from 5 to 10%. Kam-Cal$^{\text{DP}}$, Zafar$^{\text{DP}}_{\text{Fair}}$, and Kearns$^{\text{PE}}$ are exceptions; these approaches usually incur a high accuracy penalty for enforcing fairness (Figure 7(b)) and errors only further reduce accuracy by 2–4%. We note an interesting distinction between approaches that enforce demography- and error-aware fairness notions. Approaches targeting demography-aware notions cope better and their target fairness scores are typically within 5% of what they achieve in the absence of errors. These approaches repair the training data (or constrain the classifier) to meet some target demography and we hypothesize that their robustness is due to the fact that the target demography holds regardless of data errors. Thus, the drop in fairness is less severe even though accuracy is affected by

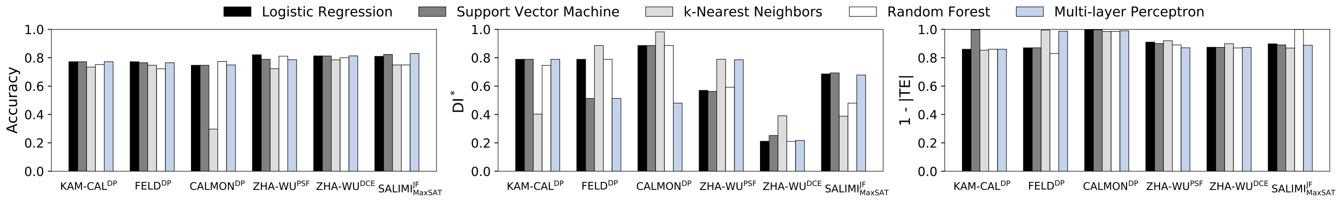


Figure 10: Sensitivity of pre-processing approaches to the choice of ML model in terms of accuracy, DI*, and TE on Adult.

corrupt data. In contrast, approaches enforcing error-aware notions are more severely impacted and we observe drops in their target fairness metric ranging from 8 to 20%. These approaches equalize error rates between the sensitive groups and heavily depend on the correctness of predictions. For instance, we observe that Calmon^{DP} and Kam-Kar^{DP} pay the least penalty in their target fairness metric even when presented with corrupt data, while Zha-Le^{EO}, Kearns^{PE}, and Thomas^{EO} all report significant drops. Finally, the changes are unpredictable for the metrics not optimized by each approach.

Key takeaway: Pre- and in-processing exhibit poor generalizability in the presence of data quality issues in the training data and fail to build models that are fair on the target population. Post-processing is more robust by design.

4.5 Sensitivity to the Underlying ML Model

All pre- and post-processing approaches need to be combined with a classifier in order to complete the ML pipeline. In this section, we study the sensitivity of pre- and post-processing approaches to the choice of ML model used for classification. We executed a new instance of each approach on the Adult dataset after pairing them with each of the following models: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN), and Multi-layer Perceptron (MLP). We implemented each classifier using Scikit-learn (version 0.22.1) and chose hyper-parameters that maximize correctness in the fairness-unaware setting (detailed in the Appendix). Figure 10 presents our results on the pre-processing approaches; we detail the rest in the Appendix.

The choice of model affects pre-processing approaches, while post-processing ones are generally less impacted. By design, post-processing approaches do not make any assumptions about the classifier that produced the predictions. Our experiments showed that their accuracy and fairness only vary slightly across different models, likely due to variation in prediction probabilities generated by each classifier. In contrast, the correctness-fairness balance of pre-processing approaches varies significantly with the choice of downstream ML model. This indicates that off-the-shelf classifier models are not always suitable for pre-processing, and hyper-parameter settings should be specific to the repaired data produced by each approach.

Key takeaway: Pre-processing is sensitive to the choice to ML model; the approaches require the hyper-parameters to be tuned separately per classifier model and in accordance to the repaired data. In contrast, post-processing is resilient to the choice of ML model and behaves similarly regardless of the model.

4.6 Other Results

In our evaluation, we further explored *data efficiency* and *stability* of all the approaches. Due to space limitations, we summarize the results here, and refer the reader to the Appendix for a full analysis. Our findings suggest that most approaches are data-efficient, and the size of the training set does not impact their accuracy and fairness significantly. In particular, most approaches produce stable results when trained on 1K data points and higher. We do not observe any significant pattern across the dimension of pre-, in-, and post-processing. Further, we find that approaches show low variance over different choices of training sets, with only a small number of outliers. Their variance tends to be higher for metrics they do not target.

Key takeaway: Approaches are generally stable and data-efficient, no stage (pre-, in-, and post-processing) appears to have an edge in these aspects.

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. Some contributions already exist in this direction (e.g., Salimi^{JF} has a parallel implementation), and improvements here are likely to lead to more impact. Notably, causality-based approaches produce sophisticated repairs, but impose a significant runtime penalty. Kam-Cal^{DP} and Feld^{DP} use simpler repairs, resulting in runtime improvement by orders of magnitude, but tend to produce poorer fairness wrt the causal metrics.

Synergy with data cleaning and repairs. Our evaluation highlighted the impact of data quality issues on the performance of pre- and in-processing techniques. Considerations of data quality are a particularly good fit for pre-processing methods, as they already focus on data repairs. Investigating repairs that combine both cleaning and fairness objectives has the potential to lead to increased robustness, which may give pre-processing approaches an edge against in-processing in practical settings.

Synergy with ML research. Our analysis noted that some in-processing techniques scale poorly 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. Further, pre-processing approaches vary in performance depending on the downstream ML model. These approaches have the potential to improve their resilience, and further investigation can explain on how to best pair these approaches with ML models.

Applicability of fairness notions and approaches. Due to the variety of notions and approaches in literature, the task of choosing the most suitable fair classification approach can be daunting. As we saw in our evaluation (Section 4.2), performance of different approaches as measured by different metrics can diverge, and it is important to follow the application requirements before attacking a problem setting with a particular method. It is similarly important to consider what fairness notions capture the nuances of and context required by the specific application.

Non-causal notions typically present the fewest computational challenges, and can be enforced efficiently (Section 4.3). However, they aim at statistical balance, often at the cost of exacerbating causal biases (Section 4.2). Causal notions provide stronger guarantees and are generally a good fit when adequate domain knowledge and computational resources are available. Enforcing multiple notions is not typically recommended, as prior literature has proved that different fairness constraints cannot be satisfied simultaneously and combining several constraints leads to a vacuous classifier [51, 67].

There are also considerable tradeoffs across the different stages of fairness enforcing mechanisms. Pre-processing presents the flexibility of being model agnostic, but there can be practical constraints to modifying training data as this may violate anti-discrimination laws [6]. Additionally, pre-processing repairs data on the assumption that model predictions will follow the ground truth. However, it cannot enforce fairness notions that balance the correctness of predictions across sensitive groups, as it cannot make assumptions on the correctness of predictions before model training. This means that notions such as equalized odds and predictive parity cannot be easily handled in the pre-processing stage. Our findings also suggest that pre-processing can pose scalability challenges with high dimensional data (Section 4.3), and can vary in performance if the downstream model is not fixed (Section 4.5). In contrast, in-processing directly modifies the learning objective, enforces a wider variety of notions, and provides better fairness guarantees. However, it is model-specific and works under the assumption that the model is replaceable, which may not be practically feasible. Similar to pre-processing, in-processing also encounters scalability issues in our experiments and their fairness guarantees may not hold if the training data contains errors (Sections 4.3–4.4). On the other hand, post-processing works on top of a trained classifier, which generally makes it more efficient and robust than pre- and in-processing. However, it often achieves poorer correctness-fairness balance, a critical component in any application. Lastly, combining multiple approaches is possible, but faces practical hurdles such as substantial penalties in correctness, runtime overhead, and required access to the entire ML pipeline.

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.

Acknowledgements: This work was supported by the NSF under grants CCF-1763423 and IIS-1943971.

REFERENCES

[1] 2015. Texas Dept. of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576U.S. (2015).

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

[3] 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

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

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

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

[17] 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.

[18] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* 4 (2020).

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

[20] 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.

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

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

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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

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

[35] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* 178, 11 (2018), 1544–1547.

[36] 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.

[37] 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.

[38] 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.

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

[40] Wen Huan, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*. 743–751.

[41] 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).

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

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

[44] 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.

[45] 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.

[46] 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.

[47] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*. 2907–2914.

[48] 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.

[49] 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.

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

[51] 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)*.

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

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

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

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

[56] 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).

[57] 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.

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

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

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

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

[62] Suvodeep Majumder, Joymallya Chakraborty, Gina R Bai, Kathryn T Stolee, and Tim Menzies. 2021. Fair Enough: Searching for Sufficient Measures of Fairness. *arXiv preprint arXiv:2110.13029* (2021).

[63] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on Causal-based Machine Learning Fairness Notions. *arXiv preprint arXiv:2010.09553* (2020).

[64] 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).

[65] Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. 2021. Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 386–400.

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

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

[68] Melissa Nobles. 2000. *Shades of citizenship: Race and the census in modern politics*. Stanford University Press.

[69] 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.

[70] Judea Pearl. 2009. *Causality*. Cambridge university press.

[71] 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.

[72] Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems*. 677–688.

[73] 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.

[74] Jonathan Rothwell. 2014. How the war on drugs damages black social mobility. *The Brookings Institution, published Sept* 30 (2014).

[75] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. 2020. Learning Certified Individually Fair Representations. In *Advances in Neural Information Processing Systems*. 7584–7596.

[76] 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.

[77] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. 2021. Automated Feature Engineering for Algorithmic Fairness. *PROCEEDINGS OF THE VLDB ENDOWMENT* 14, 9 (2021), 1694–1702.

[78] 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.

[79] 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.

[80] Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, and Gyuri Szarvas. 2018. On challenges in machine learning model management. (2018).

[81] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. 2021. JENGA-A Framework to Study the Impact of Data Errors on the Predictions of Machine Learning Models.. In *EDBT*. 529–534.

[82] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020).

[83] 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.

[84] Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*.

[85] 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.

[86] 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).

[87] 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.

[88] 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.

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

[90] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*. 1920–1953.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.*

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

[97] 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.

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

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

[100] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[101] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach.. In *IJCAI*, Vol. 16. 2718–2724.

[102] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1335–1344.

[103] 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.

APPENDIX

A DESCRIPTION OF FAIRNESS METRICS

In this section, we provide detailed discussion of the fairness metrics in Figure 4 that we choose for evaluating the fair approaches. We begin with the non-causal metrics and then continue to the causal ones in order to best highlight their rationale and differences.

A.1 Non-causal Fairness Metrics

The non-causal metrics depend entirely on empirical data and aim to establish statistical relationships between the sensitive attribute and the predictions. We start with an example that highlights two common types of discrimination typically determined from empirical data and proceed to describe how our chosen metrics operate.

Example 2. *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 11 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 ($\frac{14+6}{60} \approx 33\%$). This highlights how a group can receive an unfair advantage (or disadvantage) if the proportion of positive and negative predictions differs across groups.*
- *(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 ($\frac{TP}{TP+FN} = \frac{14}{14+2} \approx 88\%$). This indicates how predictions can disadvantage a group if the correctness of predictions differs across groups.*

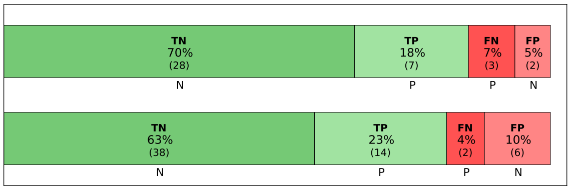


Figure 11: Prediction statistics over 100 applicants, grouped by gender: 60 male (bottom) and 40 female (top). The ground truth (positives as P and negatives as N) is indicated below each segment.

Disparate Impact (DI) is a group, non-causal, and observation level metric. It quantifies demographic parity [25], 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 2, $DI = \frac{9/40}{20/60} = 0.67$, which suggests that positive predictions are not independent 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 observation level 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 [39], 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 2, $TPRB = \frac{14}{16} - \frac{7}{10} = 0.18$ 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.

Individual Discrimination (ID) [32] is an individual, non-causal, and observation level metric. It allows us to determine both the classifier's discrimination with respect to individuals and the influence of the sensitive attribute. Specifically, ID is the fraction of tuples for which, changing the sensitive attribute causes a change in the prediction, compared to otherwise identical data points. Suppose

	$\mathbb{X}$		S	$\hat{Y}$
id	SAT	dept_choice	gender	admitted
t_1	High	Physics	Male	1
t_2	High	Mathematics	Male	0
t_3	Average	Physics	Male	1
t_4	High	Mathematics	Male	1
t_5	High	Physics	Male	1
t_6	Average	Mathematics	Male	0
t_7	high	Mathematics	Female	0
t_8	Average	Mathematics	Female	0
t_9	high	Mathematics	Female	1
t_{10}	high	Physics	Female	1
t_{11}	Average	Mathematics	Female	0
t_{12}	Average	Physics	Female	1

Figure 12: Sample data for 12 university applicants.

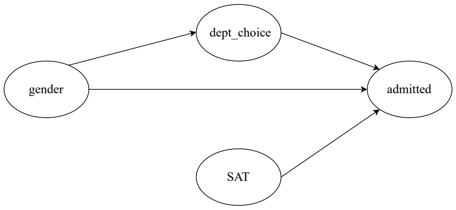


Figure 13: The graphical causal model corresponding to Example 2.

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 $ID = \frac{|Q|}{|\mathcal{D}|}$. ID lies in the range $[0, 1]$ and $ID = 0$ corresponds to perfect individual fairness as there exists no data point for which changing sensitive attribute results in a different prediction.

Example 3. *Consider 12 university applicants shown in Figure 12. To measure ID, we alter 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_7 changes from 0 to 1 when t_7's* gender *is changed from* Female *to* Male*, and that predictions do not change for any other tuples. Then, $ID = \frac{1}{12} = 0.08$, indicating that* 8% *of the applicants are discriminated because of their gender.*

The formal definition of ID requires computation over all possible data points in the domain of attributes, but practical heuristics limit interventions to smaller datasets of interest [32].

A.2 Causal Fairness Metrics

Causal fairness metrics differ from non-causal metrics in that they consider additional domain knowledge to reason about the underlying data generation process and the way changes in attributes propagate to the prediction. We first briefly discuss the structural causal model, the basic framework of causal fairness.

A structural causal model is a semantic framework that unifies the concepts of causal graphs representing causal relationships among attributes and structural equations denoting how each attribute is determined. The causal graph is typically a directed acyclic graph, where vertices represent attributes and edges represent functional relationships between these attributes. A structural equation describes the process by which an attribute changes in response to other attributes. An intervention on an attribute $X \in \mathbb{X}$, denoted by $do(X = x)$, modifies the causal model by replacing the structural equations associated with X with a constant $x \in Dom(X)$. In the causal graph, this is akin to removing all incoming edges into X and replacing with $X \leftarrow x$. The result of an intervention is a counterfactual world, where a different probability distribution is induced over the other attributes. We use $\hat{Y}_{X=x}$ to denote the potential outcome resulting from the intervention and $Pr(\hat{Y}_{X=x} = y)$ as the counterfactual distribution of predictions. These quantities can often be uniquely identified from empirical data, we refer to prior literature for a detailed discussion [70]. Most causal fairness metrics are defined in terms of interventions and counterfactuals; they can account for confounding effects present in observational data and explain the apparent discrimination using the intermediate attributes on the path from the sensitive attribute to the prediction.

Total Effect (TE) [70] is a group, causal, and intervention level metric. It intervenes on the sensitive attribute and measures discrimination as the effect of the intervention on the prediction through all causal paths.

$$TE{=}Pr(\hat{Y}_{S=1} = 1) - Pr(\hat{Y}_{S=0} = 1)$$

$TE \in [-1, 1]$ and $TE = 0$ indicates complete fairness as the causal effect of the sensitive attribute on the prediction is zero.

Example 4. *Suppose, the causal graph shown in Figure 13 represents the data generation process for Figure 12. Since the sensitive attribute (*gender*) and the prediction (*admitted*) do not share any confounders (i.e., common causes), we can easily compute TE from observational data [70]. $TE{=}Pr(\hat{Y}_{S=1} = 1) - Pr(\hat{Y}_{S=0} = 1) = Pr(\hat{Y} \mid S = 1) - Pr(\hat{Y} \mid S = 0) = \frac{4}{6} - \frac{3}{6} = 0.16\%$. This confirms that the sensitive attribute causally influences the prediction and men are overall more likely to receive to a favorable outcome.*

Natural Direct Effect (NDE) [70] is a group, causal, and intervention level metric. It computes discrimination as the expected change in prediction when the sensitive attribute changes from unprivileged to privileged, while setting other attributes to whatever value they would have attained for unprivileged individuals. Semantically, NDE measures the portion of TE that is transmitted through the direct path $S \rightarrow \hat{Y}$. Suppose, $Z \in \mathbb{X}$ is the set of attributes on the indirect paths from S to $\hat{Y}$.

$$NDE = Pr(\hat{Y}_{S=1, Z_{S=0}} = 1) - Pr(\hat{Y}_{S=0} = 1)$$

$NDE \in [-1, 1]$ and $NDE = 0$ indicates that there is no direct causal effect of the sensitive attribute on the prediction.

Example 5. *From Figure 13, the direct path corresponds to* gender⟶admitted *and the indirect path is denoted by* gender⟶ dept_choice ⟶admitted*. Using Theorem 4 of Zhang et al. [103], we can compute NDE as $\sum_{s,d} \Big\{ Pr(\hat{Y} = 1 \mid S = 1, SAT = s, dept\_choice = d) Pr(dept\_choice = d \mid S = 0) Pr(SAT = s) \Big\} - Pr(\hat{Y} = 1 \mid$*

$S = 0), \forall s \in \mathbf{Dom}(SAT), \forall d \in \mathbf{Dom}(dept\_choice)$. *Thus, NDE* $= (1 \cdot \frac{1}{2} \cdot \frac{4}{6} \cdot \frac{7}{12} + 1 \cdot \frac{2}{6} \cdot \frac{7}{12} + 1 * \cdot \frac{2}{6} \cdot \frac{5}{12}) - \frac{3}{6} = 0.01$. *This indicates that the causal influence of the sensitive attribute is minimal through the direct path and there is no significant discrimination.*

Natural Indirect Effect (NIE) [70] is a group, causal, and intervention level metric. It computes discrimination as the expected change in prediction when the sensitive attribute is unprivileged, while other attributes change to values they would have attained for privileged individuals. Semantically, *NIE* measures the portion of *TE* that is transmitted through the indirect paths $S \cdots \rightarrow Z \rightarrow \cdots \hat{Y}$.

$$NIE = Pr(\hat{Y}_{S=0, Z_{S=1}} = 1) - Pr(\hat{Y}_{S=0} = 1)$$

$NIE \in [-1, 1]$ and $NIE = 0$ indicates that no causal effect is transmitted through indirect paths from the sensitive attribute to the prediction.

Example 6. *Suppose, the Physics department has a low acceptance rate and females tend to apply there more. Using Theorem 5 of Zhang et al. [103], we can compute NIE as* $\sum_{s,d} \Big\{ Pr(\hat{Y} = 1 \mid S = 0, SAT = s, dept\_choice = d) Pr(dept\_choice = d \mid S = 1) Pr(SAT = s) \Big\} - Pr(\hat{Y} = 1 \mid S = 0), \forall s \in \mathbf{Dom}(SAT), \forall d \in \mathbf{Dom}(dept\_choice)$. *Thus,* $NIE = (1 \cdot \frac{1}{2} \cdot \frac{3}{6} \cdot \frac{7}{12} + 1 \cdot \frac{3}{6} \cdot \frac{7}{12} + 1 \cdot \frac{3}{6} \cdot \frac{5}{12}) - \frac{3}{6} = 0.14$. *This indicates that most of the causal influence is transmitted through the indirect path and females receive less favorable predictions due to their choice of department.*

B DESCRIPTION OF FAIR APPROACHES

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

B.1 Pre-processing Approaches

B.1.1 *Kam-Cal.*

Kamiran and Calders [42] 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]

B.1.2 *Feld.*

Feldman et al. [27] 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 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 $Pr(\mathcal{D}_X \mid \mathcal{D}_S = 1)$ and $Pr(\mathcal{D}_X \mid \mathcal{D}_S = 0)$. This repair produces the modified attribute $\mathcal{D}'_X$ such that $Pr(\mathcal{D}'_X \mid \mathcal{D}_S = 1) = Pr(\mathcal{D}'_X \mid \mathcal{D}_S = 0)$, and, thus, ensures that the modified attribute is independent of the sensitive attribute.

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 highly application-specific, we only report our findings for the highest level of repair ($\lambda = 1.0$). We also note that the implementation of Feld modifies both training and test data.

B.1.3 *Calmon.*

Calmon et al. [15] 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,

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

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 $Pr(\mathcal{D}'_Y \mid D_S = 0)$ and $Pr(\mathcal{D}'_Y \mid D_S = 1)$ 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, Calmon requires a distortion function to ensure that all individual values are distorted within acceptable limits.[3]

B.1.4 Zha-Wu. Zhang, Wu, and Wu [103] propose two pre-processing approaches that target *path-specific fairness* and *direct causal effect*. We refer to them as Zha-Wu$^{\text{PSF}}$ and Zha-Wu$^{\text{DCE}}$. Given training data $\mathcal{D} = [\mathcal{D}_{\mathbb{X}}, \mathcal{D}_S; \mathcal{D}_Y]$, both approaches utilize a graphical causal model to estimate the direct and indirect causal influence of the sensitive attribute on the label. Then they repair $\mathcal{D}_Y$ minimally to produce $\mathcal{D}'_Y$ such that their target fairness is achieved. Then the classifiers trained on the modified training data $\mathcal{D}' = [\mathcal{D}_{\mathbb{X}}, \mathcal{D}_S; \mathcal{D}'_Y]$ are expected to be fair, under the assumption that the distribution of the predictions made by a classifier follows the distribution of the ground truth in the training data.

Zha-Wu$^{\text{PSF}}$ enforces path specific fairness: a causal notion that ensures that the causal influence of S is not carried to Y through any direct or indirect paths. To repair $\mathcal{D}_Y$, Zha-Wu$^{\text{PSF}}$ 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$^{\text{PSF}}$ utilizes 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$^{\text{PSF}}$ 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$^{\text{PSF}}$ designs an optimization problem to repair $\mathcal{D}_Y$ such that the direct and indirect causal effects of $\mathcal{D}_S$ 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.

In contrast, Zha-Wu$^{\text{DCE}}$ aims to ensure that the direct causal effect, the influence of S that is transmitted through the direct path, is within an allowable threshold τ. It exploits the causal graph to determine a set of parents (Q) of Y, such that Q blocks all indirect paths from S to Y. To certify the presence of direct causal effect, Zha-Wu$^{\text{DCE}}$ then computes $\Delta_q = Pr(\mathcal{D}_Y = 1 \mid \mathcal{D}_S = 1, Q = q) - Pr(\mathcal{D}_Y = 1 \mid \mathcal{D}_S = 0, Q = q), \forall q \in \mathbf{Dom}\{Q\}$. Since Q blocks all indirect paths, the graphical criterion of d-separation [70] supports that $\Delta_q > 0$ occurs only if there exists some direct causal effect. Finally, Zha-Wu$^{\text{DCE}}$ modifies D_Y to ensure that $\Delta_q \leq \tau$ in each subpopulation q.

Implementation. We retrieved the source code for the approaches from the authors' website.[4] In accordance with the original papers, we set both ϵ and τ to be 0.05. The causal graphs for the datasets are constructed from prior literature (detailed in Appendix C).

B.1.5 Salimi. Salimi et al. [78] 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,[5] 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 [10] and matrix factorization [59] 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 Salimi$^{\text{JF}}_{\text{MaxSAT}}$ 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.

B.2 In-processing Approaches

Zafar. Zafar et al. [93, 95] 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

[3] https://github.com/maliha93/Fairness-Analysis-Code/blob/master/Preprocessing/Calmon/aif360/algorithms/preprocessing/optim_preproc_helpers/distortion_functions.py

[4] https://www.yongkaiwu.com/publication/

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

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.[6] 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 [83].

Implementation. We collected the source code for Zafar from the authors' public repository.[7] 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 [97] 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 [11] 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.[8]

Kearns. Kearns et al. [46] 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[9] 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 $l(\hat{Y}, Y)$:

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

While this optimization problem can be computationally hard in the worst case, Kearns computes an approximate solution by

[6] Non-convex functions are computationally harder to optimize than convex functions.

[7] https://github.com/mbilalzafar/fair-classification

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

[9] The number of subgroups must be bounded by the classifier's VC dimension [88].

solving an equivalent zero-sum game [20] 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.[8] 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. [17] 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 [38], 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.[8] 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. [85] 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:

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

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 [8] or Student's t-test [12]; 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 5), 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.

B.3 Post-processing Approaches

B.3.1 Kam-Kar. Kamiran, Karim, and others [44] 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.[10] We set all the hyperparameters following the instructions specified within the source code (more details are in the authors' repository).

B.3.2 Hardt. Hardt et al. [39] 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:

$$\begin{aligned} &\min_{g} \mathbb{E}[l(Y, \tilde{Y})] \\ \text{s.t. } &Pr(\tilde{Y} = 1 \mid S = 1, Y = y) = Pr(\tilde{Y} = 1 \mid S = 0, Y = y), \forall y \in Y, \\ \text{and } &Pr(\tilde{Y} = 1 \mid S = s, Y = y) \in [0, 1], \forall y \in Y,\ s \in S, \end{aligned}$$

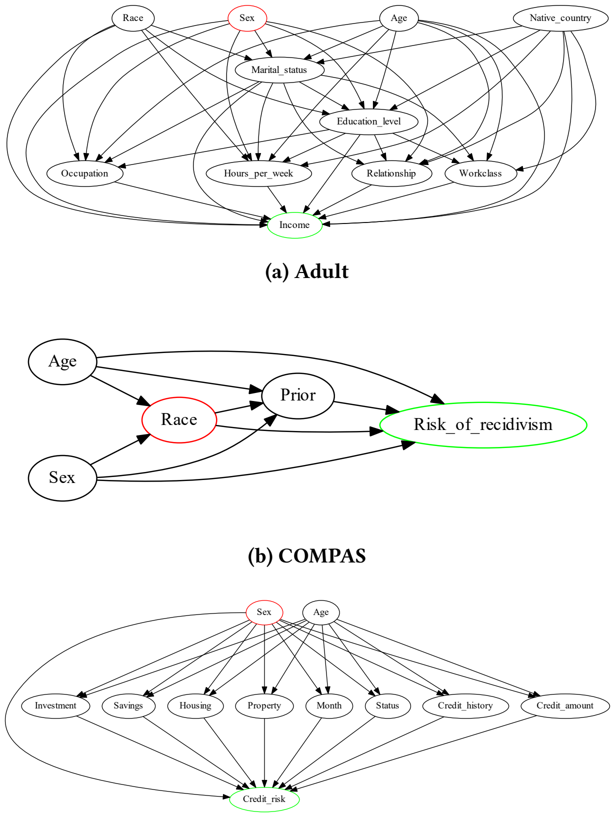
where $\mathbb{E}[l(Y, \tilde{Y})]$ is the expected loss. The solution to this linear program always provides a mapping for modifying the predictions such that equalized odds is satisfied.

Implementation. We collected the source code for Hardt from a public repository.[11]

B.3.3 Pleiss. Pleiss et al. [71] 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.



(a) Adult

(b) COMPAS

(c) German

Figure 14: The causal graph underlying each dataset in our evaluation. The nodes highlighted in red and green denote the sensitive attributes and the labels respectively.

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.[11] 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.4 Additional Approaches Under Evaluation

We evaluated 3 additional variations of 2 approaches by Madras et al. [61] and Agarwal et al. [2]. Madras et al. propose a pre-processing approach to learn a fair representation of data that assures that naively trained classifiers on it will be reasonably fair and accurate. Specifically, this approach utilizes adversarial learning by providing

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

appropriate adversarial objective functions that upper bounds the unfairness of arbitrary downstream classifiers in the limit of adversarial training. We refer to this approach as Madras^{DP} as it targets demographic parity. On the other hand, Agarwal et al. propose an in-processing approach that can enforce multiple definitions of fairness. The key mechanism is to break down fair classification to a sequence of cost-sensitive classification problems, whose solutions yield a randomized classifier with the lowest empirical error subject to the target fairness constraints. We evaluate two variations of this approach—Agarwal^{DP} and Agarwal^{EO}—that target demographic parity and equalized odds.

The aforementioned approaches are not included in the main report as Zhang et al. (Zha-Le^{EO}) utilizes adversarial learning techniques similar to Madras et al., and Celis et al. (Celis^{PP}) applies a mechanism similar to Agarwal et al. to cover a wider range of fairness notions. Figure 15 shows their correctness and fairness over Adult, COMPAS and German. Figures 20 to 24 show the results of their efficiency and scalability, robustness to data errors, sensitivity to the underlying ML model, stability, and data efficiency, respectively.

C CAUSAL GRAPHS OF ALL DATASETS

The computation of causal metrics in our evaluation require the causal structure or graphical model underlying the datasets. Figure 14 shows the causal graphs corresponding to each dataset. These causal graphs are well accepted and widely used in prior literature [66, 100, 103].

D RESULTS OF 5-FOLD CROSS VALIDATION

We cross validated each of our approach through 5-fold cross validation with 50%-20%-30% split for the train-validation-test sets. Figures 16 to 18 presents the average of each metric over all the datasets.

E COMPLETE RESULTS OF ROBUSTNESS TO DATA ERRORS

Figure 19 shows the results of our robustness experiments in terms of all correctness and fairness metrics over 3 different erroneous training set derived from COMPAS.

F COMPLETE RESULTS OF SENSITIVITY TO UNDERLYING ML MODEL

We study the sensitivity of pre- and post-processing approaches to the choice of ML model by pairing them with each of the following models: Logistic Regression (LR), Sup-port Vector Machine (SVM), Random Forest (RF), k-Nearest Neigh-bors (k-NN), and Multi-layer Perceptron (MLP). We implemented each classifier using Scikit-learn (version 0.22.1). We chose hyper-parameters that maximize correctness in the fairness-unaware setting; they are as follows:

- **LR:** `l2` regularization.
- **SVM:** `rbf` kernel with `scaled` gamma co-efficient.
- **RF:** A forest of `40` trees, each with a maximum depth of `100`.
- **k-NN:** `33` nearest neighbors.
- **MLP:** `1` hidden layer with `20` neurons, `l2` regularization with `alpha=0.01`, and `sigmoid` activation for output.

We do not mention the additional hyper-parameters associated with each classifier as they were kept at the default level. Figure 21 presents the sensitivity of all pre- and post-processing approaches to the underlying ML model over the Adult dataset.

G STABILITY OF FAIR APPROACHES

We evaluate the stability of all the approaches through a variance test on their correctness and fairness over the Adult dataset. 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 all correctness and fairness metrics in Figure 22; the results are similar over the other datasets.

Approaches are generally stable. Most approaches show low variance and have a very small number of outliers. Hardt^{EO} shows high variance in precision and F_1-score, but are stable on the other metrics. In general, approaches can exhibit slightly higher variance in the metrics they do not target.

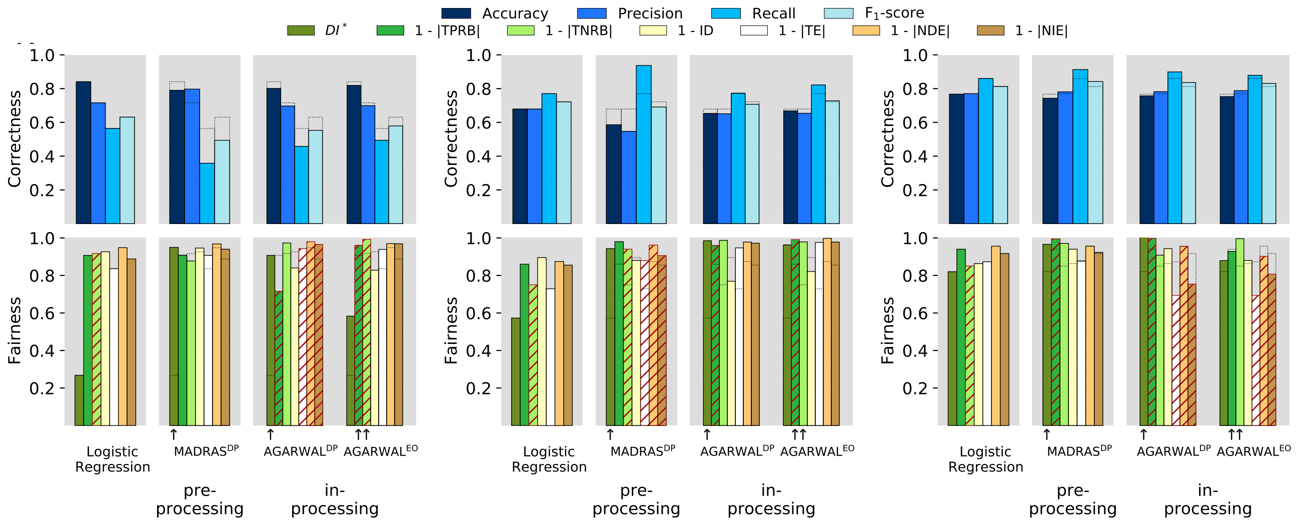
Key takeaway: All approaches generally exhibit low variance in terms of correctness and fairness over different train-test splits. High-variance behavior is rare, and there is no significant trend across the dimension of pre-, in-, and post-processing.

H DATA EFFICIENCY OF FAIR APPROACHES

In this section, we examine how the size of the training set impacts the accuracy and fairness of all approaches. We used the Adult dataset, as it is the largest, and executed a new instance of each approach ranging the size of the training set from 0.1K to 36K data points, sampled from the dataset. We do not present separate variants of the same approach unless they differ significantly in behavior. We report our findings in Figure 23.

Approaches are generally data-efficient. Most approaches we evaluate produce stable results when trained on 1K data points or higher. However, the data-efficiency can be unpredictable in metrics not optimized by an approach. Kam-Cal^{DP}, Hardt, and Pleiss^{EOP} appear to be the most data-efficient, achieving their best correctness-fairness balance with as few as 100 data points. Further, we don't observe any significant pattern across the dimension of pre-, in-, and post-processing.

Key takeaway: Most approaches are data-efficient, with the size of training data not having significant impact on their correctness-fairness balance. No stage (pre-, in-, and post-processing) appears to have an edge in this aspect.



(a) Adult (b) COMPAS (c) German

Figure 15: Correctness and fairness scores of the 3 additional fair classification approaches over (a) Adult, (b) COMPAS, and (c) German 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.

	Accuracy	Precision	Recall	F_1-score	DI*	1 - |TPRB|	1 - |TNRB|	1 - CD	1 - |TE|	1 - |NDE|	1 - |NIE|
LR	0.84	0.72	0.56	0.63	0.27	0.91	0.92	0.93	0.84	0.95	0.89
Kam-Cal^{DP}	0.77	0.64	0.21	0.32	0.79	0.96	0.99	1.00	0.86	1.00	0.86
Feld^{DP}	0.79	0.63	0.21	0.31	0.80	0.92	0.97	1.00	0.87	0.99	0.87
Calmon^{DP}	0.75	0.50	0.49	0.50	0.89	0.96	0.96	1.00	0.96	0.99	0.97
Zha-Wu^{PSF}	0.82	0.66	0.54	0.59	0.57	0.94	0.88	0.86	0.91	0.97	0.94
Zha-Wu^{DCE}	0.81	0.72	0.42	0.53	0.21	0.80	0.93	0.99	0.87	0.97	0.91
Salimi$^{JF}_{MaxSAT}$	0.81	0.72	0.54	0.62	0.61	0.90	0.93	0.96	0.90	0.97	0.94
Salimi$^{JF}_{MatFac}$	0.78	0.68	0.61	0.65	0.61	0.87	0.93	0.95	0.90	0.98	0.92
Madras^{DP}	0.79	0.79	0.35	0.49	0.95	0.91	0.88	0.95	0.91	0.97	0.94
Zafar$^{DP}_{Fair}$	0.76	0.53	0.56	0.54	0.95	0.81	0.96	1.00	0.97	0.99	0.98
Zafar$^{DP}_{Acc}$	0.82	0.70	0.49	0.58	0.74	0.82	0.99	1.00	0.98	0.99	0.99
Zafar$^{EO}_{Fair}$	0.82	0.73	0.47	0.57	0.52	0.93	0.98	1.00	0.83	0.99	0.84
Zha-Le^{EO}	0.80	0.64	0.44	0.52	0.57	0.91	1.00	0.94	0.95	0.99	0.96
Kearns^{PE}	0.82	0.79	0.36	0.50	0.26	0.89	0.96	0.96	0.80	0.96	0.84
Celis^{PP}	0.72	0.53	0.88	0.66	0.99	0.98	0.92	0.98	0.93	0.95	0.98
Thomas^{DP}	0.72	0.60	0.30	0.40	0.98	0.88	0.94	0.93	0.99	1.00	0.99
Thomas^{EO}	0.79	0.55	0.46	0.50	0.60	0.99	1.00	0.90	0.93	0.99	0.94
Agarwal^{DP}	0.80	0.69	0.45	0.55	0.90	0.71	0.97	0.84	0.94	0.98	0.96
Agarwal^{EO}	0.81	0.69	0.49	0.57	0.58	0.96	0.99	0.82	0.93	0.97	0.96
Kam-Kar^{DP}	0.76	0.52	0.78	0.62	0.95	0.85	0.92	0.93	0.96	1.00	0.96
Hardt^{EO}	0.78	0.58	0.45	0.51	0.97	0.90	0.93	0.65	0.88	0.92	0.96
Pleiss^{EOP}	0.74	0.71	0.34	0.46	0.54	0.97	0.87	0.79	0.95	0.98	0.97

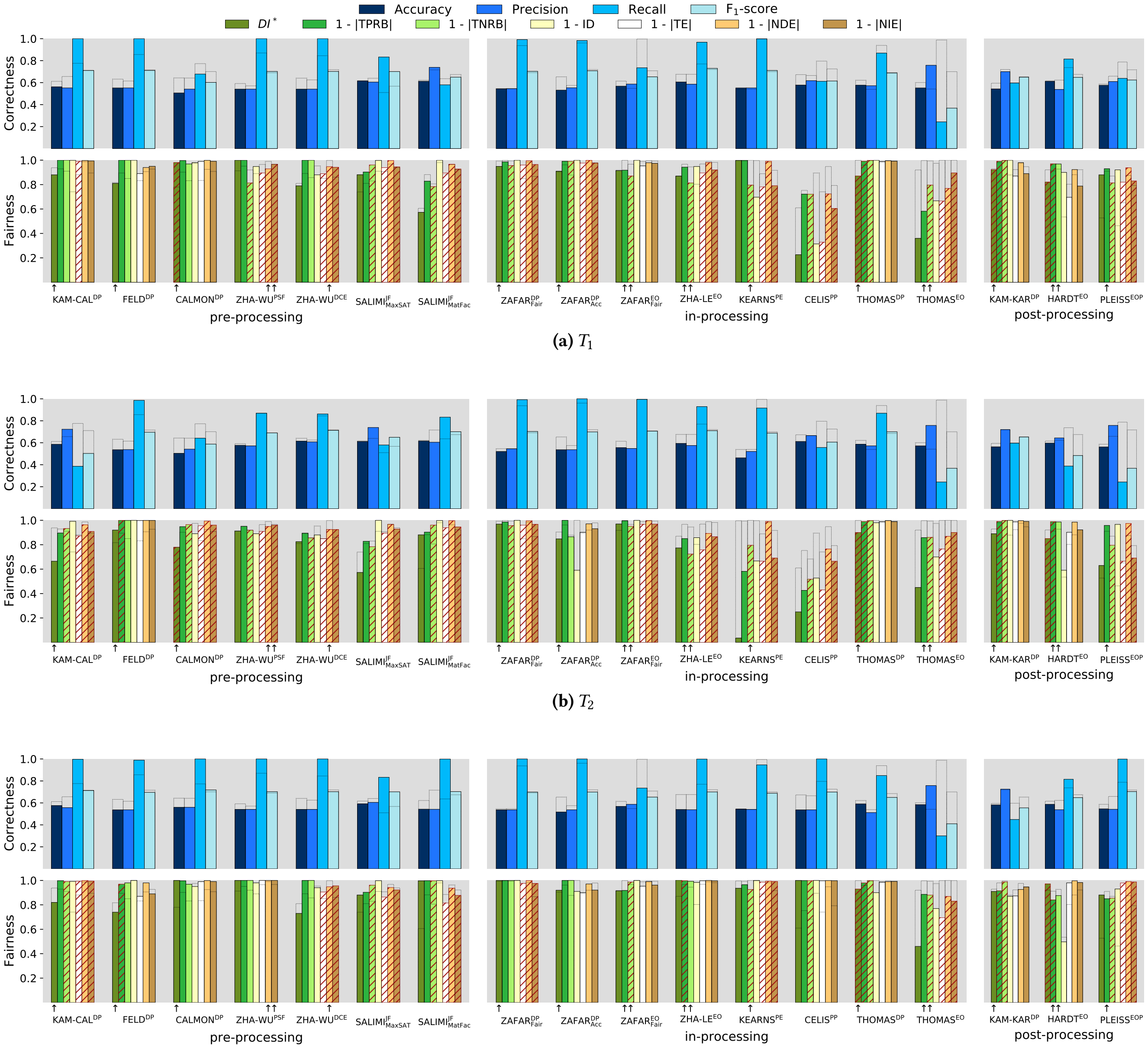
Figure 16: The average of all correctness and fairness metrics after 5-fold cross validation on Adult.

	Accuracy	Precision	Recall	F$_1$-score	DI*	1 - |TPRB|	1 - |TNRB|	1 - CD	1 - |TE|	1 - |NDE|	1 - |NIE|
LR	0.68	0.68	0.77	0.72	0.57	0.86	0.75	0.90	0.73	0.87	0.86
Kam-Cal^{DP}	0.61	0.66	0.78	0.71	0.94	0.93	0.91	0.74	0.78	0.99	0.79
Feld^{DP}	0.63	0.62	0.86	0.72	0.82	0.90	0.85	1.00	0.80	0.91	0.90
Calmon^{DP}	0.64	0.64	0.77	0.70	0.78	0.9	0.83	0.98	0.83	0.92	0.91
Zha-Wu^{PSF}	0.59	0.57	0.87	0.69	0.91	0.95	0.92	0.89	0.96	0.99	0.98
Zha-Wu^{DCE}	0.64	0.62	0.84	0.72	0.81	0.89	0.86	0.95	0.85	1.00	0.85
Salimi$^{JF}_{MaxSAT}$	0.62	0.64	0.51	0.57	0.74	0.81	0.83	0.91	0.91	0.97	0.94
Salimi$^{JF}_{MatFac}$	0.62	0.72	0.64	0.67	0.61	0.88	0.81	0.98	0.88	0.96	0.92
Madras^{DP}	0.58	0.54	0.93	0.69	0.94	0.98	0.94	0.88	0.87	0.96	0.90
Zafar$^{DP}_{Fair}$	0.55	0.55	0.94	0.69	0.99	1.00	0.99	1.00	0.85	0.98	0.87
Zafar$^{DP}_{Acc}$	0.65	0.57	0.96	0.72	0.90	0.96	0.88	1.00	0.83	0.92	0.91
Zafar$^{EO}_{Fair}$	0.61	0.55	1.00	0.71	0.92	1.00	0.97	1.00	0.99	1.00	0.99
Zha-Le^{EO}	0.68	0.68	0.77	0.72	0.87	0.97	0.94	0.80	0.87	0.99	0.88
Kearns^{PE}	0.54	0.54	1.00	0.70	1.00	0.99	1.00	1.00	0.89	0.98	0.91
Celis^{PP}	0.67	0.66	0.80	0.72	0.61	0.75	0.68	0.90	0.74	0.95	0.79
Thomas^{DP}	0.62	0.54	0.94	0.68	1.00	0.96	1.00	1.00	0.91	1.00	0.91
Thomas^{EO}	0.60	0.54	0.99	0.70	0.92	1.00	0.99	0.97	0.90	0.99	0.91
Agarwal^{DP}	0.65	0.65	0.77	0.70	0.98	0.96	0.98	0.76	0.94	0.97	0.97
Agarwal^{EO}	0.66	0.65	0.82	0.72	0.96	0.99	0.97	0.82	0.97	0.99	0.97
Kam-Kar^{DP}	0.60	0.72	0.60	0.65	0.93	0.98	1.00	0.88	0.92	0.98	0.95
Hardt^{EO}	0.62	0.62	0.74	0.68	0.92	0.91	0.93	0.53	0.80	0.88	0.92
Pleiss^{EOP}	0.59	0.66	0.79	0.72	0.53	0.91	0.87	0.46	0.65	0.88	0.78

Figure 17: The average of all correctness and fairness metrics after 5-fold cross validation on COMPAS.

	Accuracy	Precision	Recall	F$_1$-score	DI*	1 - |TPRB|	1 - |TNRB|	1 - CD	1 - |TE|	1 - |NDE|	1 - |NIE|
LR	0.77	0.77	0.86	0.81	0.82	0.94	0.85	0.86	0.87	0.96	0.92
Kam-Cal^{DP}	0.72	0.72	0.98	0.83	0.96	0.97	0.90	0.98	0.87	0.91	0.96
Feld^{DP}	0.73	0.73	0.96	0.83	0.93	0.97	0.82	1.00	0.82	0.95	0.86
Calmon^{DP}	0.76	0.78	0.91	0.84	0.99	1.00	0.97	1.00	0.87	0.95	0.92
Zha-Wu^{PSF}	0.71	0.77	0.90	0.83	0.96	1.00	0.96	0.96	0.95	1.00	0.96
Zha-Wu^{DCE}	0.74	0.77	0.89	0.83	0.70	0.80	0.65	0.79	0.90	0.98	0.92
Salimi$^{JF}_{MaxSAT}$	0.68	0.64	0.63	0.63	0.84	0.94	0.91	0.84	0.91	0.99	0.92
Salimi$^{JF}_{MatFac}$	0.70	0.63	0.67	0.65	0.85	0.97	0.89	0.89	0.90	0.95	0.95
Madras^{DP}	0.74	0.78	0.91	0.84	0.96	0.99	0.97	0.94	0.87	0.95	0.92
Zafar$^{DP}_{Fair}$	0.75	0.77	0.91	0.83	0.99	0.96	0.90	1.00	0.72	0.99	0.74
Zafar$^{DP}_{Acc}$	0.74	0.79	0.86	0.82	0.93	0.92	0.93	1.00	0.82	0.97	0.86
Zafar$^{EO}_{Fair}$	0.75	0.78	0.89	0.83	0.94	0.92	0.91	1.00	0.76	0.89	0.86
Zha-Le^{EO}	0.75	0.79	0.87	0.83	0.86	0.88	0.98	0.90	0.82	0.98	0.85
Kearns^{PE}	0.61	0.66	0.88	0.76	0.92	0.90	1.00	0.92	0.95	1.00	0.95
Celis^{PP}	0.74	0.79	0.85	0.82	0.96	0.95	0.93	0.98	0.89	0.98	0.91
Thomas^{DP}	0.73	0.71	0.88	0.79	0.99	0.98	0.94	1.00	0.87	0.94	0.93
Thomas^{EO}	0.73	0.73	0.89	0.80	0.94	0.99	0.98	0.98	0.93	0.96	0.97
Agarwal^{DP}	0.75	0.78	0.89	0.83	0.99	0.99	0.90	0.94	0.69	0.95	0.75
Agarwal^{EO}	0.75	0.78	0.87	0.83	0.87	0.92	0.99	0.88	0.69	0.90	0.80
Kam-Kar^{DP}	0.67	0.87	0.61	0.72	0.95	0.94	0.95	0.89	0.75	0.92	0.84
Hardt^{EO}	0.71	0.77	0.84	0.80	0.93	0.93	0.91	0.64	0.70	0.90	0.80
Pleiss^{EOP}	0.72	0.76	0.92	0.83	0.93	0.96	0.84	0.73	0.97	1.00	0.97

Figure 18: The average of all correctness and fairness metrics after 5-fold cross validation on German.

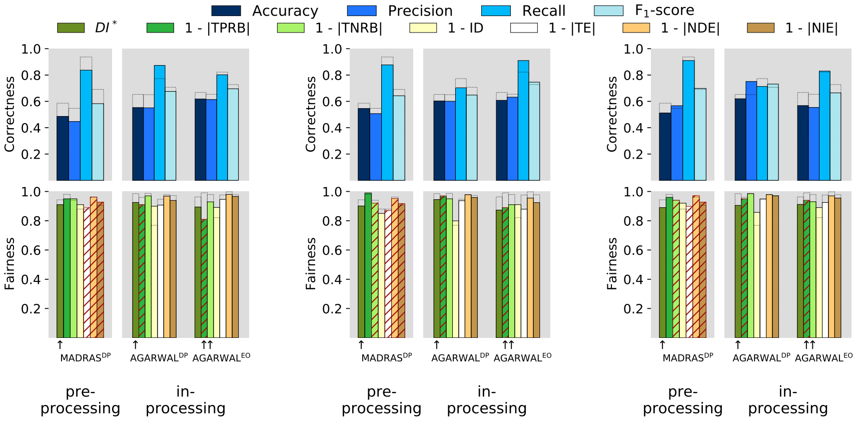


(a) T_1

(b) T_2

(c) T_3

Figure 19: The results of our robustness experiment on 3 erroneous dataset. 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 each approach on the error-free dataset are overlaid for aiding visual comparison.



(a) T_1 **(b)** T_2 **(c)** T_3

Figure 20: The results of robustness experiment on 3 erroneous dataset over 3 additional fair classifiers. 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 each approach on the error-free dataset are overlaid for aiding visual comparison.

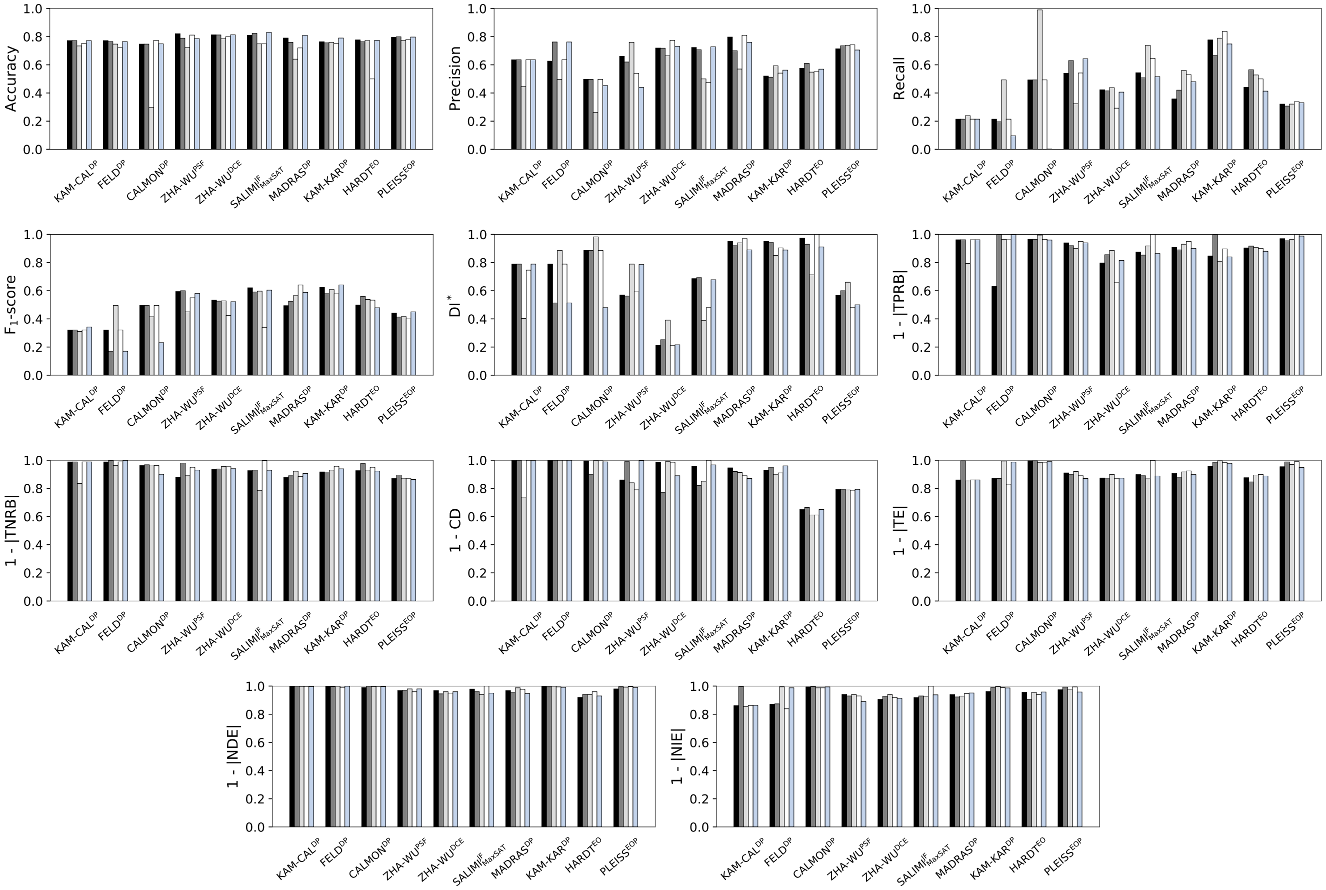


Figure 21: The sensitivity of pre- and post-processing approaches (including the additional ones) to the choice of ML model in terms of all correctness and fairness metrics on the Adult dataset.

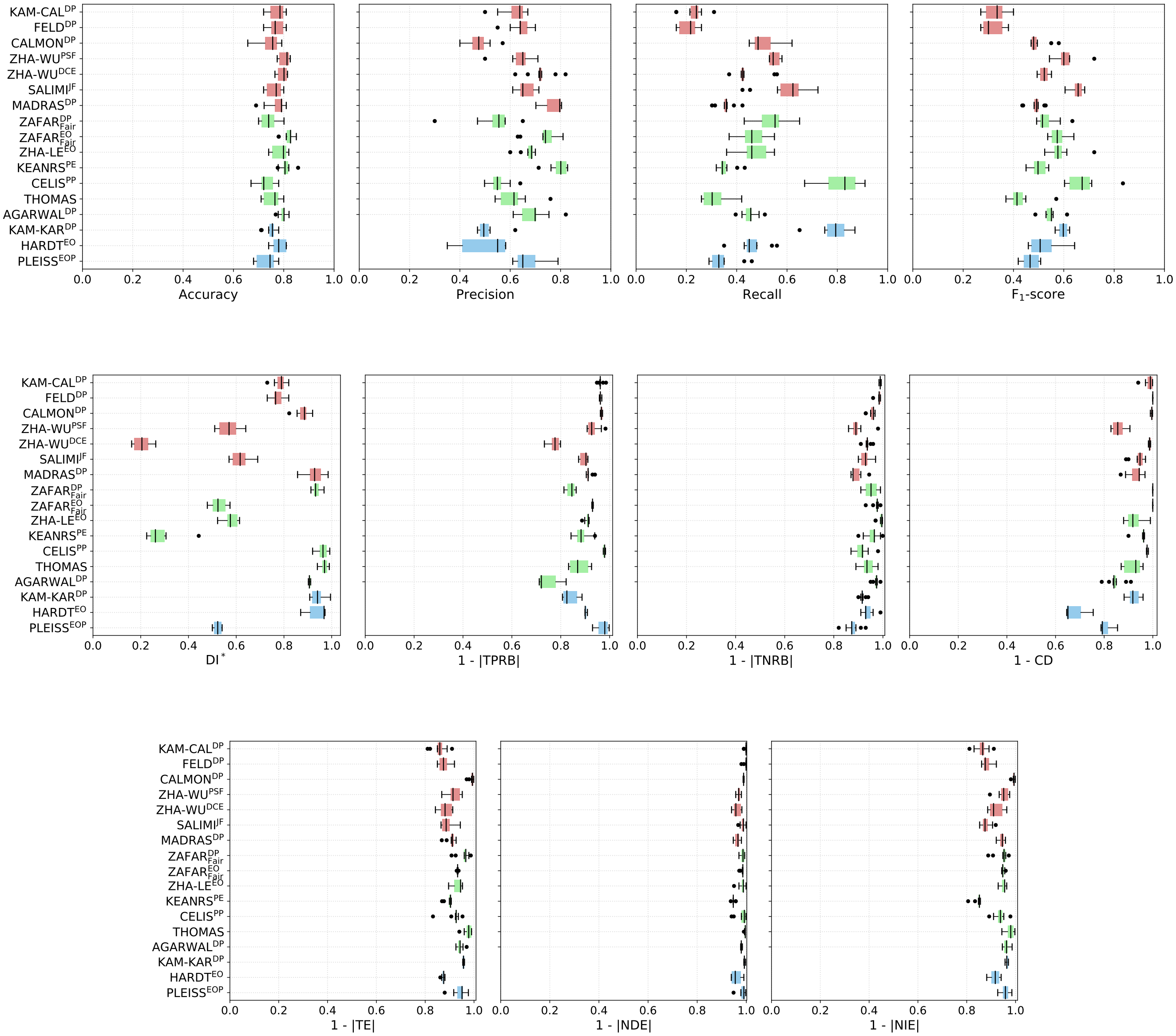
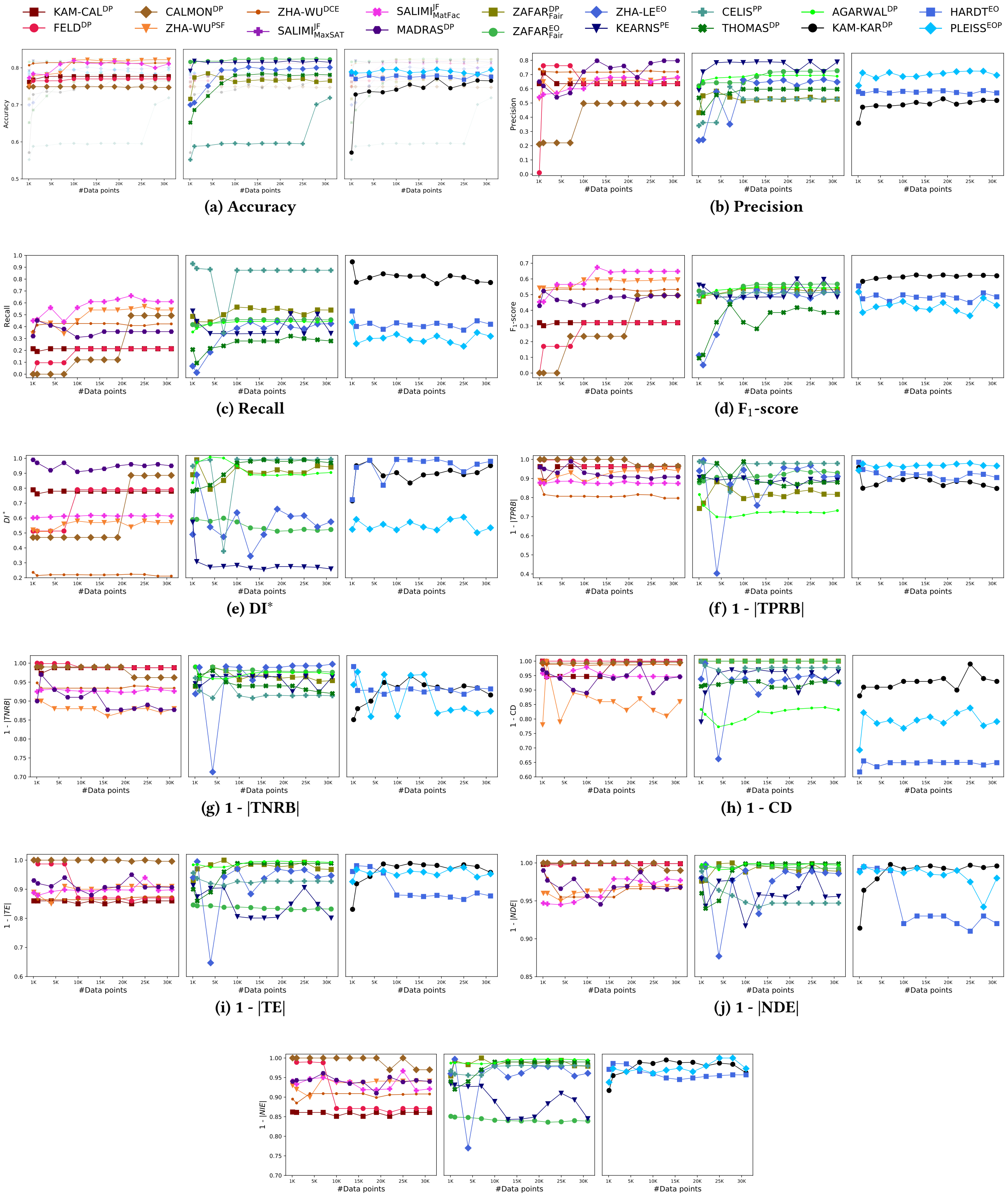


Figure 22: Variance of the fair approaches (including the additional ones) in terms of correctness and fairness metrics on arbitrary folds over Adult dataset.



(a) Accuracy

(b) Precision

(c) Recall

(d) F$_1$-score

(e) DI*

(f) 1 - |TPRB|

(g) 1 - |TNRB|

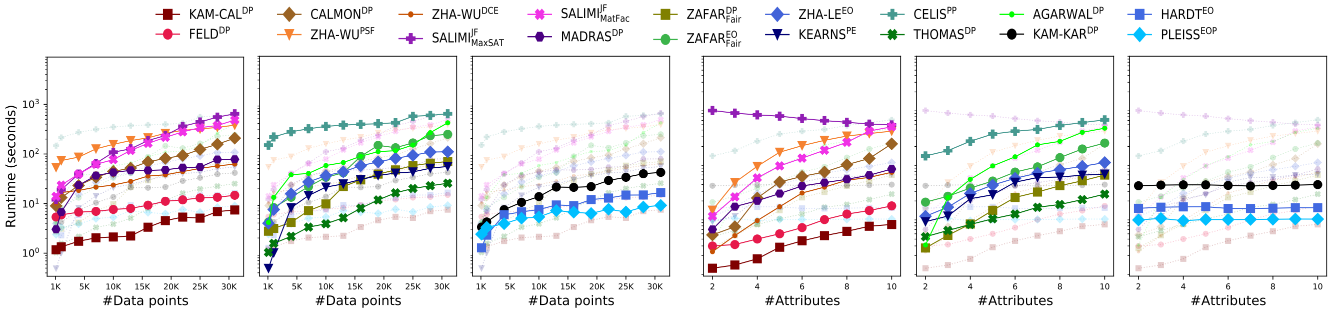
(h) 1 - CD

(i) 1 - |TE|

(j) 1 - |NDE|

(k) 1 - |NIE|

Figure 23: The data efficiency of all fair approaches (including the additional ones) on Adult in terms of all correctness and fairness metrics. Note that the y-axis is scaled differently for each metric.



(a) pre-processing **(b) in-processing** **(c) post-processing** **(d) pre-processing** **(e) in-processing** **(f) post-processing**

Figure 24: Complete results of runtime experiments on all fair approaches, including the additional ones. (a) – (c) show runtime overhead with varying data size and (d) – (f) show runtime overhead with varying number of attributes in Adult dataset. Note that the y-axis is in log scale.