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.

KEYWORDS

Empirical study; Algorithmic fairness; Classifiers

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 [39], all the way to high-stakes decisions such as mortgage approval [20], job applicant filtering [28], and pre-trial risk assessment of criminal defendants [60]. However, automated predictions are only as good as the data that drives them. As inherent biases are common in data [7], data-driven systems commonly demonstrate unfair and discriminatory behavior [10, 60, 80, 89].

Data management research has shown growing interest in the topic of fairness over applications related to ranking, data synthesis, result diversification, and others [4–6, 35, 55, 86, 94]. 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 [29, 59, 79, 80, 100], 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 [31], and achieves nearly 70% accuracy [25]. 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 [60]. 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 utilizing historical arrest data in the training set, despite certain populations being proven to be more policed than others [76].

Example 1 is not an isolated incident; other cases of classifier discrimination have pointed towards racial [10], gender [80], and other forms of bias and unfairness [89]. 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 [18, 29, 41, 80, 97], and a multitude of approaches and notions of fairness have emerged [69, 91]. 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 [69, 91], 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 [24, 53, 64] 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 [16, 29, 44, 80, 104, 105]. 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 [18, 48, 87, 95, 97, 99], and post-processing approaches, which alter the classifier predictions to ensure fairness [41, 46, 73]. 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 [17, 65, 66, 91]. Moreover, prior work on the evaluation of fair classifiers had a narrower scope than ours. Friedler et al. [33] carry out experimental analysis similar to ours by evaluating variations of 4 fair approaches over 5 fairness metrics, while Jones et al. [43] 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 [8] 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 [34, 88] 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 [16, 29, 44, 80, 104, 105], 5 in-processing [18, 48, 87, 95, 97, 99], and 3 post-processing approaches [41, 46, 73] 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 [54] and COMPAS [60]. 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)

- 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 $(\mathcal{X}, S; Y)$, where $\mathcal{X}$ 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., $\text{Dom}(S) = \{0, 1\}$, 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 $f : f(\mathcal{X}) \rightarrow \hat{Y}$, where $\hat{Y}$ denotes the predicted class label ($\text{Dom}(Y) = \text{Dom}(\hat{Y}) = \{0, 1\}$). 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 labels 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.

Metrics. We measure correctness through well-studied metrics in literature [57] (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

| Notation | Description | Metric | Definition | Range | Interpretation |
|----------|-------------|--------|------------|-------|----------------|
| $\mathcal{X}$ | A set of attributes | $TP + TN$ | $[0, 1]$ | Accuracy = 1 → completely correct |
| $\text{Dom}(\mathcal{X})$ | A single attribute $X$ and its value domain | $TP + FN$ | $[0, 1]$ | Accuracy = 0 → completely incorrect |
| $S$ | A sensitive attribute | $TP$ | $[0, 1]$ | Precision = 1 → completely correct |
| $Y, \hat{Y}$ | Attribute denoting ground-truth and predicted class label | $TP$ | $[0, 1]$ | Precision = 0 → completely incorrect |
| $\mathcal{D}$ | An annotated dataset with the schema $(\mathcal{X}, S; Y)$ | $TP$ | $[0, 1]$ | Recall = 1 → completely correct |
| $f(\mathcal{X})$ | A binary classifier | $TP$ | $[0, 1]$ | Recall = 0 → completely incorrect |
| $S_t$ | Value of the sensitive attribute $S$ for tuple $t \in \mathcal{D}$ | $\text{Precision} \times \text{Recall}$ | $[0, 1]$ | $F_1$-score = 0 → completely correct |
| $Y_t, \hat{Y}_t$ | Ground-truth and predicted class labels for tuple $t \in \mathcal{D}$ | $\text{Precision} \times \text{Recall}$ | $[0, 1]$ | $F_1$-score = 0 → completely incorrect |

Figure 1: Summary of notations.

Figure 2: List of correctness metrics used in our evaluation.
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 Pearl’s ladder of causation [72], 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 underlying 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 [72], resolving attributes that mediate the relationship between the sensitive attribute and the outcome in non-discriminatory ways [75], similarity metric between individuals [26], etc.

### Fairness metrics

While Figure 3 highlights a wide range of proposed fairness notions, Prior works [33, 64] 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 [23]. 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 [26].

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 [41], the notion they jointly measure.

Individual Discrimination (ID) [34] 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) [72], 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, 101].

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 [72]. 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; (3) post-processing approaches modify the classifier predictions. 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, under-scoring their similarities and differences. (Details are in [2]).

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 [44] 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 $Y$ reasonably approximates the ground truth $Y$, Kam-Cal argues that $Y$ 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 $D$ with a weighted sampling technique to ensure that $S$ and $Y$ are statistically independent.

Feld [29] is another approach that enforces demographic parity. It argues that demographic parity can be ensured if the marginal distribution of each attribute $X \in \mathcal{X}$ is similar across the sensitive groups in training data $D$. 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 $\mathcal{X}$ and $S$.

Calmon [16] 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 $X$ and $Y$ and without significantly distorting the underlying data distribution. It utilizes the joint distribution associated with $D$ 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 $X$ 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 [104, 105] proposes two methods that target causal notions: path-specific fairness ($Zha-Wu^{psf}$) and direct causal effect ($Zha-Wu^{dce}$). $Zha-Wu^{psf}$ enforces path-specific fairness by modifying $Y$ such that all causal influences of $S$ over $Y$ are removed. It learns a causal graph over $D$ 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, $Zha-Wu^{dce}$ 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
We choose and discuss five different in-processing approaches within the training stage and fairness is typically added as a classifier parameter that either maximizes prediction accuracy under constraints such as demographic parity [0, ∞), equalized odds [-1, 1], equalized accuracy [-1, 1], demographic parity (0, 1), and conditional accuracy [0, 1].}

3.2 In-processing

In-processing approaches are most favored by the machine learning community [18, 48, 97, 99] 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 hinge on the assumption that the model is replaceable or modifiable, which may not always be the case. We choose and discuss five different in-processing approaches and their variants, to best highlight the variety of existing techniques.

Zafar [95, 97] proposes two methods to enforce demographic parity (ZAFARDP) and equalized odds (ZAFAREOF). Both utilize tuples’ distance from the decision boundary as a proxy of Ŷ to model fairness violations, translate their corresponding fairness notion to a convex function of the classifier parameters. ZAFARDP solves the resulting constrained optimization problem to compute optimal classifier parameters that either maximize prediction accuracy under fairness constraints (ZAFARDPAC), or minimizes fairness violation under constraints on accuracy compromise (ZAFARDPFAC). ZAFAREOF only computes parameters that maximize prediction accuracy under fairness constraint [85].

Zha-Le [99] 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 D, the goal of the classifier is to maximize the accuracy of Ŷ, while the adversary attempts to correctly predict S using Ŷ and Y. Zha-Le utilizes gradient descent techniques [12] to compute the classifier’s optimal parameters such that Ŷ does not contain any information about S that the adversary can exploit.

Kearns [48] 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 [18] 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 [40] 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 notions within a single framework.

Thomas [87] 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 D and a target fairness notion, Thomas ensures that a classifier f trained on 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 predictions made by an already-trained classifier. Their benefit is that they do not require classifier retraining.
According to the stage of the classifier pipeline where fairness-enhancing mechanism is applied, these approaches are divided into three categories: pre-processing, in-processing, and 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.

Figure 5: List of fair approaches, fairness notions they support, and high-level descriptions of the mechanisms they apply to ensure fairness.

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** [46] 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 $\tilde{Y}$ for tuples in that region until the probability of positive outcome is similar across sensitive groups, i.e., demographic parity is achieved.

**HARDT** [41] enforces equalized odds through modifying the predictions $\tilde{Y}$. Given access to $Y$ and $S$ from the training data $D$, HARDT learns the parameters of a new mapping $g : g(\tilde{Y}, S) \rightarrow \tilde{Y}$ to replace $\tilde{Y}$ such that TPR and TNR are equalized across the sensitive groups. The new mapping is learned by solving a linear program.

**PLEISS** [73] 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 $\tilde{Y}$ 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 more fair approaches (not in Figure 5) in the Appendix. While other fair approaches exist, some are incorporated in the ones we evaluate [14, 15, 45], while others are empirically inferior [47], offer weaker guarantees [3, 74], do not offer a practical solution [71, 92], or do not apply to the classification setting [38, 58, 62, 81]. Some make strong assumptions about the problem setting [22, 51, 56, 67, 68, 78, 101, 102], or require additional information [26, 59, 77, 98], 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 corrupted data; however, post-processing is generally more robust than pre-processing and in-processing.

Next we provide details on our experimental settings: evaluated approaches, their implementation details, evaluation metrics, and the datasets. Then we 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 (LR) as the classifier. This is in line with the evaluations of the original papers as they all use LR. Moreover, to contrast all fair approaches against a fairness-unaware approach, we trained an unconstrained LR classifier 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 [8] (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 [84] to compute causal quantities. We implemented the evaluation scripts in Python 3.6 [27].

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 $DI^* = \min(DI_{\text{min}}^*, \frac{DI}{DI_{\text{min}}})$, which ensures that low fairness with respect to $DI$ ($DI \to 0$ and $DI \to \infty$) is mapped to low values for $DI^*$. Further, we report $1 - |TPRB|$, $1 - |TNRB|$, $1 - ID$, $1 - |TE|$: 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 [33, 43, 66].

Adult [54] 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 [60] 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 [25]. 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 [36] 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.

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

| Dataset | Size (MB) | $|D|$ | $|X|$ | $S$ | Sensitive groups | Target task |
|---------|-----------|------|------|-----|--------------------|-------------|
| Adult   | 3.70      | 45,222 | 9  | Sex | Female | Male | Income ≥ $50K |
| COMPAS  | 8.18      | 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-attributebased discrimination. We provide the target prediction tasks in the rightmost column.
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\textsuperscript{DP} and CALMON\textsuperscript{DP}) demonstrate a much larger accuracy drop than the approaches that target equalized odds (such as ZAFAR\textsuperscript{FAIR}, ZHA-LE\textsuperscript{EO}, and KEARNS\textsuperscript{EO}). ZAFAR\textsuperscript{DP}, 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, HARD\textsuperscript{EO} and PLEISS\textsuperscript{EO}, 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.
We noted a significant difference in the proportions of TEing 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.

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 [23]. 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 [30] 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).

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 [38], ZHA-WU [33], and SALMI [39], 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.

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 than pre- and in-processing. This is because post-processing operates on less similarity 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—e.g., FELD [36], ZAFAR [35], and ZAFAR [34]—trivially satisfy ID by discarding the sensitive attribute while training, even though they do not.

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.

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.

Post-processing approaches tend to violate individual level fairness.

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.

Key takeaway: 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 is sub-optimal with finite training data [92], post-processing approaches typically achieve inferior correctness-fairness balance compared to other approaches. In all the datasets, post-processing 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

We now study the runtime behavior of all approaches, to investigate their efficiency gap and highlight the need for scalability considerations. While some approaches can benefit from optimizations, such as the use of GPUs, producing these optimizations is beyond our scope. We do not present separate variants of the same approach unless they differ significantly in behavior. We compute the total runtime of each approach as pre-processing time (if any) + training time + post-processing time (if any). We subtract from all methods the runtime of LR, so that what we report is the overhead each approach introduces over the fairness-unaware method.

Our first experiment investigates the efficiency and scalability of the 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\textsuperscript{DP} and THOMAS\textsuperscript{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\textsubscript{PSF}, Zha-Wu\textsubscript{DCE}, and SALIMI\textsubscript{FP}—incurred significantly higher runtimes compared to other pre-processing approaches. In fact, both variations of SALIMI\textsubscript{FP} are NP-hard in nature. Simply, discovering causal associations from data is more complex than non-causal associations. CALMON\textsuperscript{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\textsubscript{PSF}, Zha-Wu\textsubscript{DCE}, SALIMI\textsubscript{FP}, and CALMON\textsuperscript{DP}) and the rest (KAM-CAL\textsuperscript{DP} and Feld\textsuperscript{DP}). In fact, KAM-CAL\textsuperscript{DP} and Feld\textsuperscript{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 \( \text{SALMI}^{\text{MASAT}} \) 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 \( \text{SALMI}^{\text{MASAT}} \) increases rapidly with fewer attributes, resulting in higher runtimes in those settings.

**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 [82, 83]. Furthermore, data quality issues are highly correlated with sensitive attributes in many domains like healthcare and immigration [19, 70]. 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 [37].

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 \( F_1 \)-score, and 5-10% decrease in the fairness metrics the approaches optimize.

Pre- and in-processing 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\(^{\text{IP}}\), Zafar\(^{\text{IP}}\), and Kearsn\(^{\text{IP}}\) are exceptions: they usually incur high accuracy penalties 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 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\(^{\text{IP}}\) and Kam-Kar\(^{\text{IP}}\) pay the least penalty in their target fairness metric even when presented with corrupt data, while Zha-Le\(^{\text{IP}}\), Kearsn\(^{\text{IP}}\), and Thomas\(^{\text{IP}}\) 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 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.
We discuss these general guidelines here.

The goal of our work has been to bring some clarity to the vast and variance over different choices of training sets, with only a small likely have impact in improving in-processing approaches in that overlooked in ML research, and data management contributions can particularly good fit for pre-processing methods, as they already focus on problems with high dimensional data, and can vary in performance if the downstream 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. 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 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, IIS-1949371, and 2112606. We thank Sainyam Galhotra for useful discussions.
Specifically, we observe two ways females are being discriminated:

- **Discrimination-1**: The fraction of females predicted as highly-qualified (positive) is \( \frac{7}{14} = 50\% \) which is significantly lower than the fraction of males predicted as highly-qualified (78% in total). 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 \( \frac{7}{14} = 50\% \) which is significantly lower than that of the males (78% in total). This indicates how predictions can disadvantage a group if the correctness of predictions differs across groups.

Disparate Impact (DI) is a group, non-causal, and observation level metric. It quantifies demographic parity [26], 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** 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, \( DI = \frac{30}{28} = 0.67 \), 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.

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

Both \( \text{TPRB} \) and \( \text{TNRB} \) lie in the range \([-1, 1]\). These two metrics, together, measure equalized odds [41], 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 \( \text{TPRB} \) and \( \text{TNRB} \) 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, \( \text{TPRB} = \frac{14}{16} - \frac{8}{28} = 0.18 \) and \( \text{TNRB} = \frac{10}{28} - \frac{16}{27} = -0.07 \). The high positive value of \( \text{TPRB} \) indicates **Discrimination-2**: the TPR of females is much lower than males.

**Individual Discrimination (ID)** [34] 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 that \( Q \) is the set of such tuples, defined as \( Q = \{ a \in D \mid \exists b : X_a = X_b \land S_a \neq S_b \land \hat{Y}_a \neq \hat{Y}_b \} \); then \( ID = \frac{|Q|}{|D|} \). \( ID \) 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 \( t \) changes from 0 to 1 when \( t \)’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 [34].

### 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 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 \( X \in \mathcal{X} \), denoted by \( \text{do}(X = x) \), modifies the causal model by replacing the structural equation of \( X \) with an equation \( X = x \).

The **Natural Direct Effect (NDE)** [72] 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.

\[
\text{NDE} = \Pr(\hat{Y}_{S = 1} = 1) - \Pr(\hat{Y}_{S = 0} = 1)
\]

\( \text{NDE} \in [-1, 1] \) and \( \text{NDE} = 0 \) 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 \( \text{TE} \) from observational data [72]. \( \text{TE} = \Pr(\hat{Y}_{\text{S = 1}} = 1) - \Pr(\hat{Y}_{\text{S = 0}} = 1) = \Pr(\hat{Y} \mid S = 1) - \Pr(\hat{Y} \mid S = 0) = \frac{5}{6} - \frac{7}{8} = 0.16 \%. \) This confirms that the sensitive attribute causally influences the prediction and men are overall more likely to receive a favorable outcome.
that is transmitted through the direct path $S \to \hat{Y}$. Suppose, $Z \in X$ is the set of attributes on the indirect paths from $S$ to $\hat{Y}$.

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

$NDE \in [-1, 1]$ 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. [105], we can compute $NDE$ as $\sum_{s,a} Pr(\hat{Y} = 1 \mid S = 1, SAT = s, dept\_choice = d)Pr(dept\_choice = d \mid S = 0)Pr(SAT = s) - Pr(\hat{Y} = 1 \mid S = 0)$, $\forall s \in \text{Dom}(SAT)$, $\forall d \in \text{Dom}(dept\_choice)$. Thus, $NDE = (1 \cdot \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{1}{2}) - \frac{1}{2} = 0.01$. 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)** [72] 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 due to their choice of department.

$$\text{NIE}(S) = \text{Pr}(Y = 1 \mid S = 1) - \text{Pr}(Y = 1 \mid S = 0)$$

$\text{NIE}(S) \in [-1, 1]$ and $\text{NIE}(S) = 0$ 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. [105], we can compute $\text{NIE}(S) = (1 \cdot \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{1}{2}) - \frac{1}{2} = 0.14$. 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 [44] 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 $Y$ 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 $Pr_{exp}(S = s \land Y = y)$ should be sufficiently close to their observed joint probability $Pr_{obs}(S = s \land Y = y)$. These probabilities (over $\mathcal{D}$) are computed using the following formulas:

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

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

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:

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

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.1

**B.1.2 Feld.** Feldman et al. [29] 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 $\mathcal{X}$ 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 $\mathcal{D} = \{\mathcal{D}_X, \mathcal{D}_S, \mathcal{D}_Y\}$ with the schema $(\mathcal{X}, S, Y)$, Feld produces a modified dataset $\mathcal{D}' = \{\mathcal{D}'_X, \mathcal{D}'_S, \mathcal{D}'_Y\}$ where the marginal distributions of each attribute is similar across the privileged and the unprivileged groups. Feld repairs values of each individual attribute separately to equalize the marginal distributions 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 $Pr(\mathcal{D}_X \mid \mathcal{D}_S = 0)$ and $Pr(\mathcal{D}_X \mid \mathcal{D}_S = 1)$.

Repeating the repair process for all attributes produces the modified $\mathcal{D}'_X$ 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.1 As the preferred value of $\lambda$ 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. [16] 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 $\mathcal{X}$ and $Y$ such that the dependency between $Y$ and $S$ is reduced, without significantly distorting the

1https://github.com/Trusted-AI/AIF360/tree/master/aif360/algorithms/preprocessing
To certify the presence of direct causal effect, the authors propose to compute the joint distribution $D' = \{D'_S, D'_D, D'_Y\}$. To do so, the new joint distribution is constructed following the following constraints that must be satisfied: (1) the difference between $P_r(D'_Y | D_S = 0)$ and $P_r(D'_Y | D_S = 1)$ is below an allowable threshold, (2) the new joint distribution is sufficiently close to the original one, and (3) no attribute value in $D'_X$ is substantially distorted to compute $D'_X$. The authors then formulate a convex optimization problem that searches for the optimal new joint distribution subject to the constraints. The result is a new dataset $D'$ where the influence of fairness is repaired in $D$. To repair $D'$, the authors first verify if $D_Y$ violates path-specific fairness. Specifically, $D'_Y$ is a direct or indirect cause of $D_Y$ when intervening on $D_S$ changes the expectations of $D_Y$. $Zha-Wu$ utilizes the graphical causal model and estimates the effect of intervening on $D_S$ as the expected difference in $D_Y$ when $D_S$ changes from privileged to unprivileged. Instead of measuring causal association through all paths between $D_S$ and $D_Y$ in the causal graph, $Zha-Wu$ can measure this association through specific paths if desired. Path-specific fairness is violated if the expected difference in $D_Y$ is above some threshold $\epsilon$. Next, $Zha-Wu$ designs an optimization problem to repair $D_Y$ such that the direct and indirect causal effects of $D_S$ are removed, and the causal model is minimally altered. The modified training dataset $D'$ is then used to train classifiers that enforce path-specific fairness.

In contrast, $Zha-Wu$ assumes that $Y$ is likely to be fair if a classifier is trained on data $D$ where ground truth $Y$ satisfies the target fairness notion. To that end, it expresses justifiable fairness as an integrity constraint and repairs $D$ to ensure that the constraint holds on the repaired training data $D'$. Unlike Kam-Cal, $Salimi$ does not modify the attributes and only repairs $D$ 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 (B). A contains the attributes that are allowed to influence or have causal associations with prediction $Y$, while $I$ contains the rest of the attributes. Given A and I, justifiable fairness holds in $D$ if $Y$ is independent of I conditioned on A. If the probability distribution associated with $D$ is uniform, this integrity constraint can be checked through the following multi-valued dependency: $D = \Pi_Y(D) \Rightarrow \Pi_I(D)$.

The goal of $Salimi$ is then to minimally repair $D$ to form a new training dataset $D'$, such that the multi-valued dependency is satisfied. $Salimi$ leverages techniques from maximum satisfiability [11] and matrix factorization [61] to compute the minimal repair of $D$ that produces the optimal $D'$ 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 $Calmon$ from the AI Fairness library. $Calmon$ requires a distortion function to ensure that all individual values are distorted within acceptable limits.[2]

### B.1.4 Zha-Wu

Zhang, Wu, and Wu [105] propose two pre-processing approaches that target path-specific fairness and direct causal effect. They refer to them as $Zha-Wu^{psf}$ and $Zha-Wu^{dce}$. Given training data $D = \{D_Y, D_S, D_X\}$, 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 $D_Y$ minimally to produce $D'_Y$ such that their target fairness is achieved. Then the classifiers trained on the modified training data $D' = \{D_S, D_X; D'_Y\}$ 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.

$Zha-Wu^{psf}$ 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 $D_Y$, $Zha-Wu^{psf}$ first verifies if $D_Y$ violates path-specific fairness. Specifically, $D_S$ is a direct or indirect cause of $D_Y$ when intervening on $D_S$ changes the expectations of $D_Y$. $Zha-Wu^{psf}$ utilizes the graphical causal model and estimates the effect of intervening on $D_S$ as the expected difference in $D_Y$ when $D_S$ changes from privileged to unprivileged. Instead of measuring causal association through all paths between $D_S$ and $D_Y$ in the causal graph, $Zha-Wu$ can measure this association through specific paths if desired. Path-specific fairness is violated if the expected difference in $D_Y$ is above some threshold $\epsilon$. Next, $Zha-Wu^{psf}$ designs an optimization problem to repair $D_Y$ such that the direct and indirect causal effects of $D_S$ are removed, and the causal model is minimally altered. The modified training dataset $D'$ is then used to train classifiers that enforce path-specific fairness.

In contrast, $Zha-Wu^{dce}$ aims to ensure that the direct causal effect, the influence of $S$ that is transmitted through the direct path, is within an allowable threshold $\tau$. 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, $Zha-Wu^{dce}$ then computes $\Delta_Y = Pr(D_Y = 1 | D_S = 1, Q = q) - Pr(D_Y = 1 | D_S = 0, Q = q)$, $\forall q \in \text{Dom}(Q)$. Since $Q$ blocks all indirect paths, the graphical criterion of dis-separation [72] supports that $\Delta_Y > 0$ occurs only if there exists some direct causal effect. Finally, $Zha-Wu^{dce}$ modifies $D_Y$ to ensure that $\Delta_Y \leq \tau$ in each subpopulation $q$.

Implementation. We retrieved the source code for the approaches from the authors’ website. In accordance with the original papers, we set both $\epsilon$ and $\tau$ 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. [80] 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 $Y$, 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 $Y$ is likely to be fair if a classifier is trained on data $D$ where ground truth $Y$ satisfies the target fairness notion. To that end, it expresses justifiable fairness as an integrity constraint and repairs $D$ to ensure that the constraint holds on the repaired training data $D'$. Unlike Kam-Cal, $Salimi$ does not modify the attributes and only repairs $D$ 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 (B). A contains the attributes that are allowed to influence or have causal associations with prediction $Y$, while $I$ contains the rest of the attributes. Given A and I, justifiable fairness holds in $D$ if $Y$ is independent of I conditioned on A. If the probability distribution associated with $D$ is uniform, this integrity constraint can be checked through the following multi-valued dependency: $D = \Pi_Y(D) \Rightarrow \Pi_I(D)$.

The goal of $Salimi$ is then to minimally repair $D$ to form a new training dataset $D'$, such that the multi-valued dependency is satisfied. $Salimi$ leverages techniques from maximum satisfiability [11] and matrix factorization [61] to compute the minimal repair of $D$ that produces the optimal $D'$ 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 $Salimi^{psf}$ 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. [95, 97] propose two in-processing approaches to enforce demographic parity and equalized odds. We refer to them as $Zafar^{dpar}$ and $Zafar^{eod}$, respectively. Both of these approaches translate their corresponding fairness notion to a convex function of the classifier parameters, and compute the optimal parameters that minimize prediction errors while satisfying the notion.

To compute the optimal fair classifier, $Zafar$ first formulates the learning process as a constrained optimization problem. Given the training data $D$, 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 $\theta$, 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 $d_p(X_t) \geq 0$, where $d_p(X_t)$ denotes the signed distance. $Zafar$ does not explicitly use $S$ to determine the prediction, rather they utilize $S$ to define the fairness constraint only.

$Zafar^{dpar}$ introduces a proxy constraint for demographic parity, as directly including the notion as a constraint leads to non-convexity in the loss function. $Zafar^{eod}$ utilizes $d_q$ as a proxy for $Y$ 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 $\theta$, it can be used define the proxy constraint for demographic parity. Formally, covariance is computed as: $\text{cov} = \frac{1}{m} \sum_{t \in T} (S_t - \bar{S}) d_q(X_t)$, where

---

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

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

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

6Non-convex functions are computationally harder to optimize than convex functions.
Kearns begins by formulating the learning process and constraint for the target fairness notion. Let \( f : (X, S) \rightarrow Y \) be a classifier learned over training data \( D = (X, S, Y) \). Moreover, let \( G \) be the set of the subgroups for which fairness must be ensured. Each \( g \in G \) indicates a subgroup such that \( g(X, S_i) = 1 \) means tuple \( t \) belongs to subgroup \( g \). If predictive equality is the target notion, a group function \( \beta(g) = \Pr(Y = 1 | S = 1) - \Pr(Y = 1 | S = 0) \) 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 \( \gamma \) 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 \( I(\tilde{Y}, Y) \):

\[
\min_{f} I(\tilde{Y}, Y)
\]

subject to \( \alpha(g) \beta(g) \leq \gamma, \forall g \in G \).

While this optimization problem can be computationally hard in the worst case, Kearns computes an approximate solution by solving an equivalent zero-sum game [21] 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. 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. [18] 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 \( D = (X, S, Y) \) 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 \( g_i = (X, S = i, Y) \subseteq D \). For each group in \( G \), Celis then defines \( q_i(f) \) that is a linear function or quotient of linear functions of \( \Pr(Y = 1 | g_i, \varepsilon_i) \), where \( \varepsilon_i \) 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:

\[
\min_{f} \max_{g \in G} q_i(f) \geq \tau, \quad \forall i \in [0, 1]
\]

where \( \tau \) is 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:

\[
\min_{f} \Pr(f(X) \neq Y)
\]

subject to \( \min_{g \in G} q_i(f) \geq \max_{g \in G} q_i(f) \geq \tau, \forall i \in [0, 1] \).

To solve the above problem efficiently, Celis solves its dual instead using Lagrange duality [40], which produces an approximately fair classifier. This fair classifier can only guarantee \( \min_{g \in G} q_i(f) \geq \tau - \max_{g \in G} q_i(f) = - \kappa \), where \( \kappa \) denotes additional error from estimating the probability distribution of data from samples in \( D \).

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

**THOMAS**. Thomas et al. [87] 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 $D$ and a target fairness notion, THOMAS ensures that a classifier $f$ trained on $D$ 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:

$$\underset{f}{\text{argmax}} \ L(f)$$

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

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: $D_1$ and $D_2$. Thomas then uses gradient descent to compute a candidate solution that maximizes the objective function on $D_1$. Using $D_2$, 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 [9] or Student’s t-test [13]; 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 [46] 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 $f : f(X, S) \rightarrow \hat{Y}$ be a classifier and $Pr(\hat{Y} | X, S)$ be the prediction confidence. KAM-KAR defines a critical region around the decision boundary where the prediction confidence is below a threshold $\theta$, i.e., $\max(Pr(\hat{Y} = 1 | X, S), Pr(\hat{Y} = 0 | X, S)) < \theta$. Here, $\theta$ 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. We set all the hyper-parameters following the instructions specified within the source code (more details are in the authors’ repository).

#### B.3.2 HARDT

Hardt et al. [41] 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 $q : q(Y, S) \rightarrow \tilde{Y}$ 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 $q$ must satisfy the following condition: $Pr(\hat{Y} = 1 | S = 1, Y = y) = Pr(\tilde{Y} = 1 | S = 0, Y = y), \forall y \in Y$. Given any standard loss function $l : l(Y, \tilde{Y}) \rightarrow \mathbb{R}$ that quantifies the cost of incorrect predictions, HARDT solves the following linear program to obtain the optimal mapping:

$$\min \mathbb{E}[l(Y, \hat{Y})]$$

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

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

**Implementation.** We collected the source code for HARDT from a public repository.10

#### B.3.3 PLEISS

Pleiss et al. [73] 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 $D$, is reliable and calibrated, i.e., $Pr(Y = 1 | \hat{Y} = y) = y, \forall y \in Y$. 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 $\alpha$ depends on the difference between $C_0$ and $C_1$. Then $\hat{Y}_t$ is replaced with $\tilde{Y}_t$, such that $\tilde{Y}_t = 1$ with probability proportional to the fraction of positive tuples in the favored group. This modification technique decreases the classifier’s performance for the favored group while maintaining classifier calibration, and approximately satisfies the target fairness notion. Pleiss et al. acknowledge that their approach satisfies group-level fairness while intentionally violating individual-level fairness due to randomness in predictions.

**Implementation.** We collected the source code for PLEISS from the authors’ public repository.10 We use equal opportunity as the fairness notion, since minimizing the difference in terms of favorable outcomes—i.e., equal TPR across the sensitive groups—is more appropriate as the fairness goal in the context of our datasets. Further, a weighted combination of equal opportunity and predictive equality led to very poor performance in terms of correctness in most cases.
We evaluated 3 additional variations of 2 approaches by Madras et al. [63] and Agarwal et al. [3]. Madras et al. propose a pre-processing approach to learn a fair representation of data that assures that naïvely trained classifiers on it will be reasonably fair and accurate. Specifically, this approach utilizes adversarial learning by providing 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$^{99}$ 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$^{89}$ and AGARWAL$^{99}$—that target demographic parity and equalized odds.

The aforementioned approaches are not included in the main report as Zhang et al. (ZHA-LSTM$^{99}$) utilizes adversarial learning techniques similar to Madras et al., and Celis et al. (Celis$^{97}$) 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 [68, 102, 105].

### 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), 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). We chose hyper-parameters that maximize correctness in the fairness-unaware setting; they are as follows:

- **LR**: 12 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 28 neurons, 12 regularization with $\text{alpha}=0.01$, 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. Hard$^{99}$ shows high variance in precision and F1-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\textsuperscript{DP}, Hardt, and Pleiss\textsuperscript{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.
Table 1: The average of all correctness and fairness metrics after 5-fold cross validation on Adult.

| Method      | Accuracy | Precision | Recall | F1-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|
| KM-CalDP    | 0.61     | 0.66      | 0.78   | 0.71     | 0.94| 0.93| 0.91| 0.74| 0.78| 0.99| 0.79|
| FeldDP      | 0.63     | 0.62      | 0.86   | 0.72     | 0.82| 0.90| 0.85| 1.00| 0.80| 0.91| 0.90|
| CalmonDP    | 0.64     | 0.64      | 0.77   | 0.70     | 0.78| 0.9| 0.83| 0.98| 0.83| 0.92| 0.91|
| Zha-WuPSG   | 0.59     | 0.57      | 0.87   | 0.69     | 0.91| 0.95| 0.92| 0.89| 0.96| 0.99| 0.98|
| Zha-WuDE    | 0.64     | 0.62      | 0.84   | 0.72     | 0.81| 0.89| 0.86| 0.95| 0.85| 1.00| 0.85|
| SalimMatFac | 0.62     | 0.64      | 0.51   | 0.57     | 0.74| 0.81| 0.83| 0.91| 0.91| 0.97| 0.94|
| MadrasDP    | 0.58     | 0.54      | 0.93   | 0.69     | 0.94| 0.98| 0.94| 0.88| 0.87| 0.96| 0.90|
| ZafarDP     | 0.55     | 0.55      | 0.94   | 0.69     | 0.99| 1.00| 0.99| 1.00| 0.85| 0.98| 0.87|
| ZafarFairDP | 0.65     | 0.57      | 0.96   | 0.72     | 0.90| 0.96| 0.88| 1.00| 0.83| 0.92| 0.91|
| ZafarED     | 0.61     | 0.55      | 1.00   | 0.71     | 0.92| 1.00| 0.97| 1.00| 0.99| 1.00| 0.99|
| Zha-LEED    | 0.68     | 0.68      | 0.77   | 0.72     | 0.87| 0.97| 0.94| 0.80| 0.87| 0.99| 0.88|
| KearnsDP    | 0.54     | 0.54      | 1.00   | 0.70     | 0.99| 1.00| 1.00| 0.89| 0.98| 0.91| 0.91|
| CellSSP     | 0.67     | 0.66      | 0.80   | 0.72     | 0.61| 0.75| 0.68| 0.90| 0.74| 0.95| 0.79|
| ThomasDP    | 0.62     | 0.54      | 0.94   | 0.68     | 1.00| 0.96| 1.00| 1.00| 0.91| 1.00| 0.91|
| ThomasTO    | 0.60     | 0.54      | 0.99   | 0.70     | 0.92| 1.00| 0.99| 0.97| 0.90| 0.99| 0.91|
| Agarwalo     | 0.65     | 0.65      | 0.77   | 0.70     | 0.98| 0.96| 0.98| 0.76| 0.94| 0.97| 0.97|
| AgarwaloED  | 0.66     | 0.65      | 0.82   | 0.72     | 0.96| 0.99| 0.97| 0.82| 0.97| 0.99| 0.97|
| Kam-KarDP   | 0.60     | 0.70      | 0.60   | 0.65     | 0.93| 0.98| 1.00| 0.88| 0.92| 0.98| 0.95|
| HardDE      | 0.62     | 0.62      | 0.74   | 0.68     | 0.92| 0.91| 0.93| 0.53| 0.80| 0.88| 0.92|
| PleissED    | 0.59     | 0.66      | 0.79   | 0.72     | 0.53| 0.91| 0.87| 0.46| 0.65| 0.88| 0.78|

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

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$^{DP}$ | 0.68      | 0.64   | 0.63        | 0.63| 0.84           | 0.94           | 0.91     | 0.84     | 0.91     | 0.99     | 0.92     |
| Salimi$^{HAT}$ | 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$^{DP}$  | 0.75      | 0.77   | 0.91        | 0.83| 0.99           | 0.96           | 0.90     | 1.00     | 0.72     | 0.99     | 0.74     |
| Zafar$^{DCC}$ | 0.74      | 0.79   | 0.86        | 0.82| 0.93           | 0.92           | 0.93     | 1.00     | 0.82     | 0.97     | 0.86     |
| Zafar$^{EO}$  | 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$^{EO}$ | 0.61      | 0.66   | 0.88        | 0.76| 0.92           | 0.90           | 1.00     | 0.92     | 0.95     | 1.00     | 0.95     |
| Celis$^{EO}$  | 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      | 0.61   | 0.72        | 0.95| 0.94           | 0.95           | 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$^{DP}$ | 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.
Figure 19: The results of our robustness experiment on 3 erroneous 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 each approach on the error-free dataset are overlaid for aiding visual comparison.
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.
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.
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.