Towards Profiling Fair Classification Approaches

Maliha Tashfia Islam **Zafeiria Moumoulidou**
College of College of Information & Computer Sciences
University of Massachusetts Amherst
{mtislam, zmoumoulidou}@cs.umass.edu

Abstract

Classification is a fundamental supervised-learning method that is frequently used in high-stake decision making. Nonetheless, it has been noticed that classification systems often discriminate against historically underrepresented groups. As a result, during the last couple of years the machine learning community designed various mechanisms that aim to provide *fairness* guarantees. In this project we select six fair classification approaches and evaluate their profile in terms of *correctness*, *fairness*, *efficiency*, and *scalability* metrics. Further, in our evaluation we use two real-world datasets and contrast with four fundamental *fairness-unaware* classification models: Logistic Regression, Support Vector Machines, Decision Trees and Neural Networks. To the best of our knowledge, the novelty of this study is that it offers a thorough comparison among linear and non-linear models using a diverse set of fair classification mechanisms, correctness and performance metrics. Our findings show that fair approaches generally tradeoff some correctness for fairness. All fair approaches enhance fairness according to the fairness metrics we chose to evaluate on, but there is no single best approach that does best across all correctness and fairness metrics. Further, our findings show that different approaches have different level of efficiency, and we specifically identify the approaches that are the most scalable with increasing number of data points and attributes.

1 Introduction

Data-driven predictions and classification systems shape our perception of the world and ultimately affect our decisions in a plethora of scenarios: from the places we visit, the news we read, and the movies we watch, to who we find more suitable for a job position or who we identify as prone to criminal activity. However, systems behave unexpectedly and make mistakes that often disproportionately affect various demographic groups: e.g., Amazon's hiring tool was found to systematically identify male candidates as more qualified for software engineering positions than female ones [11], the misclassification error rate of COMPAS—a pre-trial risk assessment tool—was nearly twice as high for black defendants, who were erroneously classified as high-risk for recidivism, compared to white defendants [18], and commercial face recognition systems have higher accuracy in white and male faces [4].

Consequently, the idea of fair classification has emerged in the recent literature along two primary directions: *fairness notions* that dictate what fairness in classification should mean, and *fairness mechanisms* that entail how a target notion of fairness should be enforced. This has led to an explosion of fair classification approaches that apply different mechanisms to enforce different notions. Fairness notions either aim for equal treatment among *individuals* or *groups*. The notion of individual fairness suggests that any two people with similar characteristics should be treated similarly regardless of their sensitive information (e.g., race, gender etc.), while group fairness suggests that demographic groups defined with respect to a sensitive attribute receive similar treatment. Moreover, at a high level, the currently known fair approaches can be grouped into three main categories

according to their mechanisms: (1) *pre-processing* approaches that modify the training data to repair biases, (2) *in-processing* approaches that enforce fairness constraints in the learning algorithms, and (3) *post-processing* approaches that alter the predictions of classification models to ensure fairness in their output.

Although a plethora of fair classification approaches are now available in the literature, there is a need to study and profile the fair approaches (or classifiers) in order to understand their strengths and weaknesses. Fair approaches differ in correctness and fairness due to the different notions of fairness they enforce, while the mechanism itself affects the runtime performance. The tradeoffs between the different approaches need to be studied more exhaustively. To that end, our goal in this work is to study six state-of-the-art fair approaches and empirically analyze their tradeoffs. We summarize our contributions, methodology and experimental results as follows:

- We provide an overview of the following six *fair* approaches: (1) the *pre-processing* methods of Feldman et al. [8] and Calmon et al. [5], (2) the *in-processing* method of Zafar et al. [25] and Celis et al. [6], and (3) the *post-processing* method of Hardt et al. [12] and Pleiss et al. [21].
- In order to highlight the difference in behavior, We contrast the chosen fair approaches with four widely-used *fairness-unaware* models: Logistic Regression, Support Vector Machines (SVMs), Decision Trees, and Neural Networks.
- In order to evaluate the *correctness* and *fairness* of the approaches, we select three correctness metrics—accuracy, precision, recall—and three fairness metrics—statistical parity difference, true positive balance rate, false positive balance rate—in our experiment. Further, we use two real-world datasets, Adult [16] and Compas [18], that represent different phenomena of discrimination.
- At a high level, the experimental approach we selected is the following: we first train a fairness-unaware classifier. Then using the fairness mechanisms we study in this project we build a fair classifier and evaluate how its behavior, in terms of correctness and fairness, compares to the corresponding unfair model.
- Our findings demonstrate that typically fair approaches trade accuracy for fairness. Further, we observe that all of the mechanisms are typically able to reduce bias and discrimination accroding to the fairness metrics we evaluate on. However, we observe there is no single mechanism that outperforms all the other mechanisms across all metrics.
- We also study the *efficiency* and *scalability* of fair approaches. In particular, we analyze their runtime performance as the number of data points increases. We repeat the experiment for increasing number of attributes. Our findings show that pre- and in-processing approaches demonstrate varying degree of efficiency and scalability, but post-processing approaches are overall the most efficient and scalable.

2 Related Work

Fair classification approaches. In recent years there have been a variety of fair approaches proposed in the literature. As it is not possible to evaluate all fair approaches within the scope of our project, we highlight some noteworthy fair approaches that we did not include in our evaluation.

In terms of pre-processing approaches, Kamiran et al. [14] is a work that has been very popular over the years. It is a reweighing technique that samples the tuples in the data in order to equalize the proportion of positive/negative labels across sensitive groups (e.g., people of different races). Nevertheless, the pre-processing approaches we choose for our evaluation (Section 4) empirically out-perform this approach [13]. Further, there have been recent works in terms of causality based approaches [17, 27] that explore the cause and effect relationship between the sensitive attribute and the predicted outcome. For example, Kusner et al. [17] describe a counter-factual world where fair classifiers are trained only on derived attributes that are not causally influenced by the sensitive attribute. However, all causality based mechanisms require significant domain knowledge to properly design the causal model underlying the data.

In terms of in-processing, much of the work involving fair classification has targeted this direction. Zhang et al. [26] utilize techniques from adversarial learning to train a fair classifier and adversary simultaneously, where fairness is ensured if the adversary cannot guess the sensitive attribute based on the predictions made by the classifier, i.e., predictions are similar regardless of the sensitive group. On the other hand, other works [1, 22] define constrained optimization techniques that

accommodate multiple fairness constraints/notions within a single framework. One of our evaluated approaches (Celis et al. [6]) is also a general framework with stronger theoretical guarantees than the aforementioned approaches.

In terms of post-processing, it is the least popular dimension for applying fairness enhancing techniques as it is generally less flexible than pre- or in-processing. Kamiran et al. [15] propose to modify the predictions of the tuples that are very close to the decision boundary—where the classifier is least confident about its predictions—in order to ensure the proportion of favorable outcomes is equal across the sensitive groups. However, a recent study by Woodworth et al. [24] proves the sub-optimality of post-hoc corrections and discusses a (theoretical) first step towards designing nearly-optimal post-processing approaches.

Experimental analysis of fair approaches. While there has been a variety of work towards developing fair classification approaches, prior work in terms of profiling or benchmarking these approaches has been very limited. The most relevant work that coincides with ours is an experimental evaluation by Friedler et al. [9]. This work compares variations of 4 fair approaches over 5 fairness metrics, and primarily explores issues such as the impact of data pre-processing, correlation between different measures of fairness, etc. However, this work is limited to pre- and in-processing fair approaches and does not consider scalability or efficiency issues. Further, we added more recent fairness-unaware ML models (e.g., neural networks) in our comparative analysis. Another related work close to ours is the AI-Fairness 360 toolkit by IBM [3], which is an extensible framework that includes a variety of fair approaches and metrics for testing. Nonetheless, it is not designed for comparative analysis of approaches on an equal footing, rather used for exploring different approaches individually. Other works [23, 10] provide general frameworks to evaluate approaches on a specific fairness metric but are not extendable for evaluating multiple metrics. Lastly, there are surveys that discuss fair approaches and metrics available in the literature [2, 19], but do not empirically evaluate them.

3 Novelty Statement

The aim of this project is to select some of the frequently used fairness-unaware classification models and fair classification approaches in the literature. We empirically compare them in terms of their performance aspect, such as correctness, fairness, efficiency, and scalability. As we discussed in detail in Section 2, prior work in terms of profiling fair approaches has been limited. Prior works either do not consider performance issues like efficiency, or do not empirically evaluate/analyze fair approaches across different metrics. Our project has the following novelties: **(1)** it demonstrates an empirical study of fair approaches that well-represent the state-of-the-art, **(2)** it includes approaches from all dimensions (pre-, in-, and post-processing) unlike previous work [9], and lastly, **(3)** it presents efficiency and scalability experiments that were not previously explored.

4 Methodology

In this section we define the correctness and fairness metrics according to which we evaluate the behavior of the fairness approaches we study in this project. Further, we provide an overview of the fair approaches under our evaluation in order to give a high level understanding of their mechanisms.

4.1 Metrics

We first introduce some notation that we use throughout this section. Specifically, we define $y = \{0, 1\}$ to be the true label for a data sample; we refer to $y = 1$ as the *positive* and to $y = 0$ as the *negative* class respectively. In our evaluation we interpret $y = 1$ to be the *favorable* outcome of the classification task.

Further, we define $\hat{y}$ to be the predicted label for a sample. We also model fairness in terms of a binary sensitive attribute $S = \{0, 1\}$; we define the $S = 1$ group to be *privileged* and the $S = 0$ group to be *unprivileged*. Throughout this paper, we refer to the privileged and the unprivileged groups together as *sensitive groups*.

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positives (TP)	False Positives (FP)
$\hat{y} = 0$	False Negatives (FN)	True Negatives (TN)

y : true label, $\hat{y}$: predicted label

Figure 1: Confusion Matrix for a binary classifier.

4.1.1 Correctness Metrics

In Figure 1 we define the confusion matrix for a binary classifier and introduce the notions of true positives/negatives and false positives/negatives. We now define the following metrics:

Accuracy. The accuracy of a classifier is defined as the number of samples that are correctly classified, either as positively or negatively labeled, and is equal to:

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$

The values of this metric lie within the $[0, 1]$ range; larger accuracy values indicate higher correctness for a classifier.

Precision. The precision of a classifier is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

Recall. The recall of a classifier is:

$$Recall = \frac{TP}{TP + FN}$$

Similarly to accuracy the values of precision and recall metrics also lie within the $[0, 1]$ range. Higher values indicate higher correctness. Intuitively, precision models the success rate in identifying samples that belong in the positive class. On the other hand, recall measures the percentage of samples in the positive class that were correctly classified.

4.1.2 Fairness Metrics

Below we describe the fairness metrics we use in our experimental evaluation.

Statistical Parity Difference. Statistical parity difference measures if each demographic group receives the same proportion of favorable outcomes, i.e., is the probability of receiving favorable outcomes same for everyone irrespective of whichever sensitive group they belong to. To that end, this metric is defined as follows:

$$SPD = p(\hat{y} = 1 \mid s = 1) - p(\hat{y} = 1 \mid s = 0)$$

The values of this metric lie in the $[-1, 1]$ range. $SPD = 0$ indicates a completely fair classifier, i.e., each group receives equal proportion of positive/favorable outcomes. On the other hand, as $|SPD|$ moves closer to 1 the classifier becomes more unfair. Notice that this metric does not depend on the true label of the samples.

True Positive Balance Rate (TPBR). TPBR measures a classifier's ability to correctly identify members of the positive class, irrespective of whichever sensitive group they belong to. The aim is that the true positive rate should be similar in the privileged and the unprivileged group, so that the classifier does not discriminate (or favor) against one group. This metric is defined as:

$$TPBR = p(\hat{y} = 1 \mid y = 1, s = 1) - p(\hat{y} = 1 \mid y = 1, s = 0)$$

False Positive Balance Rate (FPBR). Similarly to TPBR, FBPR measures if the false positive rate is similar across the sensitive groups. It is defined as:

$$FPBR = p(\hat{y} = 1 \mid y = 0, s = 1) - p(\hat{y} = 1 \mid y = 0, s = 0)$$

Both TPBR and FBPR take values in the $[-1, 1]$ range. The closer $\mid$ TPBR $\mid$ (or $\mid$ FBPR $\mid$) is to zero the fairer the classifier, it indicates there is little difference in success rates between the sensitive groups and no group has unfair advantage.

4.2 Fair approaches

The approaches we evaluate can be categorized using the dimension of pre-, in-, or post-processing. Hence, we provide a brief overview of the approaches using this categorization.

4.2.1 Pre-processing approaches

Pre- processing approaches are motivated from the fact that machine learning techniques are data-driven and the predictions of a classifier reflect the trend and biases of the training data. All pre-processing approaches work under the assumption that the distribution of predictions will reasonably follow the training labels. These approaches modify the data before training to remove biases, which subsequently ensures that the predictions of a learned classifier satisfy the targeted notion of fairness. The main advantage of pre-processing is that it is model-agnostic, which allows flexibility in choosing the classification algorithm based on the application requirements. However, since pre-processing happens *before* training and does not actually have access to the predictions, the approaches often do not come with provable guarantees for fairness. We describe the approaches we use in this project:

Feldman et al. [8] (FELDMAN) propose a pre-processing approach that enforces fairness as *statistical parity*, which requires the proportion of positive predictions to be similar across the sensitive groups. This approach argues that if a classifier is learned on training data where the attributes are independent of the sensitive attribute S, then the classifier is also likely to make predictions that are independent of S. Thus, the proportion of positive predictions will be similar across sensitive groups. To this end, Feldman et al. [8] modify each attribute in the training data, and ensure that the marginal distribution of each individual attribute is indistinguishable across the sensitive groups.

Calmon et al. [5] (CALMON) define a pre-processing approach that also enforces fairness as *statistical parity*. This approach defines a new joint distribution and modifies the attributes in the training data as well as the training labels. It modifies the data with three goals in mind: (1) the amount of statistical dependence between training labels and the sensitive attribute is reduced, so that a classifier trained on this data satisfies statistical parity, (2) the joint distribution of the transformed data is close to the original distribution, and (3) individual attributes are not substantially distorted.

4.2.2 In-processing approaches

In-processing approaches are most favored by the machine learning community 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, avoids any extra overhead of pre- or post-processing. However, in-processing techniques are model-specific and require re-implementation of the learning algorithms to include the fairness constraints. This hinges on the assumption that the model is replaceable or modifiable, which may not always be the case.

Zafar et al. [25] (ZAFAR) propose an in-processing approach that enforces fairness as *statistical parity*. The actual definition of statistical parity (Section 4.1.2) is not directly used as a constraint, as it is not a convex function of the classifier parameters. Instead, Zafar et al. [25] utilize tuples' distance from the decision boundary as a proxy of $\hat{y}$ and models a proxy constraint of statistical parity that is also a convex function of the classifier's parameters. Then this approach solves the resulting constrained optimization problem that maximizes prediction accuracy under fairness constraints.

Celis et al. [6] (CELIS) present an in-processing approach that accommodates a large group of fairness notions or metrics, unlike the approaches we described so far. To that end, this approach develops a general framework that solves constrained optimization problems under the fairness constraint that expresses the targeted fairness notion. It relaxes all fairness notions (or metrics) to a linear function of the classifier parameters and designs the corresponding optimization problem to minimize prediction error under fairness constraints. In our experiments, we use a variation of this approach that uses a fairness notion called *predictive parity*[1], which targets to equalize the false discovery rate among sensitive groups.

4.2.3 Post-processing approaches

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

Hardt et al. [12] (HARDT) propose a post-processing that enforces the notion of *equalized odds*, which seeks to equalize the TPR and TNR across the sensitive groups (i.e., minimizes TPBR and FPBR). This approach learns a new predictor derived from the predictions $\hat{y}$ and the sensitive attributes S. Essentially, it solves a linear program to find probabilities with which to change predictions so as to enforce equalized odds.

Pleiss et al. [21] (PLEISS) define a post-processing approach that enforces *predictive equality*—a notion that equalizes FPR across the sensitive groups (i.e., minimizes FPBR)—while also ensuring that the classifier predictions remain calibrated. To achieve this, it modifies prediction $\hat{y}$ for a random subset of tuples within the sensitive group with lower FPR in order to equalize across the groups.

5 Experimental Evaluation

In this section we describe the all the experimental settings such as the datasets, hyper-parameters for each model, etc. We further discuss the experimental methodology according to which we produced the results that we present at the end of this section. All the necessary documentation and source code for our experiments are publicly available[2].

5.1 Datasets

We select two real-world datasets that are frequently used in the fairness literature because of their imbalanced nature and bias. We describe them below:

Adult dataset [16]. This dataset is extracted from the US Census and consists of $n = 45,220$ tuples (after removing tuples with missing attributes) that report income-related information about individuals. The number of available features is 14 and include age, sex, race, capital-gain, occupation information among others; in total there are 6 continuous and 8 categorical attributes. The classification task in this dataset is to predict whether the salary for a given individual exceeds $\$50$K. Specifically, favorable or positive labels equal to 1 ($y = 1$) indicate that an individual earns $\geq \$50$K. Adult is highly discriminatory towards females as females in this dataset have significantly lower proportion of positive labels, i.e., females are less likely to have receive income $\geq \$50$K. Hence, in our evaluation we choose sex to be the sensitive attribute and females constitute the unprivileged group.

Compas dataset [18]. The compas dataset from ProPublica consists of $n = 5,278$ tuples that contain information about the profile of defendants. There are 8 available features including race, age, length of imprisonment etc. The prediction task is to identify whether a defendant re-offends within two years of the initial arrest; positive labels $(y = 1)$ indicate that an individual did not re-offend. In our evaluation we choose race to be the sensitive attribute as Compas demonstrates the most race-based discrimination: the number of re-offenders is much higher than all other races. We chose African-Americans as the unprivileged group and all other races as privileged.

[1] The source code from the original authors only had this variation as publicly available.

[2] https://github.com/imoumoulidou/Project_689

Method	Hyper-parameters	Method	Hyper-parameters
Decision Tree (DT)	$d = 10$	Decision Tree (DT)	$d = 3$
FELDMAN (DT)	$d = 10$	FELDMAN (DT)	$d = 3$
CALMON (DT)	$d = 10$	CALMON (DT)	$d = 5$
Support Vector Machine (SVM)	$C = 10, \gamma = 0.001$	Support Vector Machine (SVM)	$C = 1, \gamma = 1$
FELDMAN (SVM)	$C = 10, \gamma = 1$	FELDMAN (SVM)	$C = 10, \gamma = 1$
CALMON (SVM)	$C = 10, \gamma = 0.01$	CALMON (SVM)	$C = 10, \gamma = 0.01$
Adult Dataset		Compas Dataset	

Figure 2: Hyper-parameters for the implemented classifiers. (Left): Adult dataset, (Right): Compas Dataset.

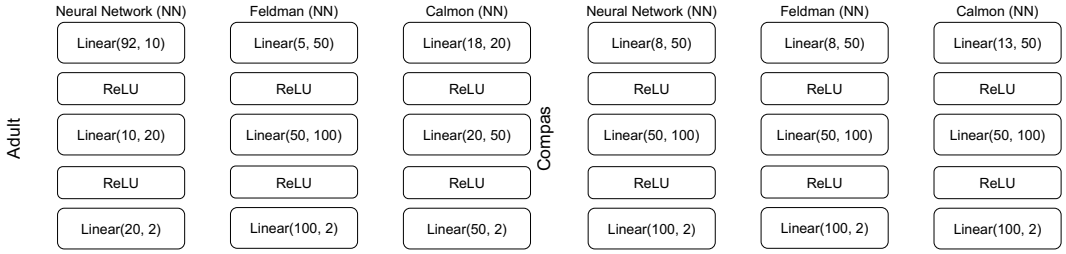

Figure 3: The model architecture of the neural networks we designed. We use softmax as the ouput layer and cross entropy as the optimization objective.

5.2 Correctness and Fairness Evaluation: Methodology

5.2.1 Fairness-unaware Machine Learning Models

In this section we describe the process we followed to build the four fairness-unaware machine learning models on each of the datasets we use in our experimental evaluation. Specifically, we use Logistic Regression (LR), Decision Trees (DTs), Support Vector Machines (SVMs) and Neural Networks (NNs). Hence, We have a total of **8** fairness-unaware models across the datasets (as seen in Figure 4).

Dataset Preparation. We pre-processed the Adult and Compas dataset to binarize categorical features and standarize the data to zero mean and unit variance. Further, we split the data into a $70-30\%$ train-test fold for LR, DTs, and SVMs models. For Neural Networks we used a $70 - 15 - 15\%$ split for training, validation and test data respectively.

Training Process and Hyper-parameter Tuning. In this section we describe the methodology we followed to train and select the best machine learning for the two classification tasks we studied. For the implementation of LR, SVMs and DTs we use the scikit-learn library [20], and for NNs we use PyTorch.

Decision Trees. We perform a 5-fold validation to select the best value for the depth value. We experimented with the following depth values $d = \{3, 5, 10, 15, 20, 25, 30, 40\}$.

Support Vector Machines. We select rbf as the kernel for the SVM models and perform a grid search over C, γ parameters using 5-fold validation to select their best combination. C parameter controls the regularization strength, which inversely proportional to the value of C. Further, γ controls how much points far from the decision boundary affect the model. We tried the following combinations: $C = \{10^1, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and $\gamma = \{1, 10^{-1}, 10^{-2}, 10^{-3}\}$.

Neural Networks. For each classification task, we experimented with various model architectures. Specifically, we experimented with the number of layers in the MLP and the number of hidden units. We used cross entropy as the objective to be minimized and ReLU as the activation function. Further we used Adam as the optimizer and experimented with various learning rates including

Method	Accuracy	Precision	Recall	| SPD|	| TPBR |	| FPBR |	Method	Accuracy	Precision	Recall	| SPD|	| TPBR |	| FPBR |
Logistic Regression (LR)	0.85	0.74	0.60	0.19	0.08	0.01	Decision Tree (DT)	0.84	0.77	0.52	0.13	0.04	0.053
FELDMAN (LR)	0.8	0.68	0.35	0.037	0.017	0.05	FELDMAN (DT)	0.8	0.65	0.39	0.01	0.04	0.04
CALMON (LR)	0.78	0.61	0.41	0.12	0.072	0.2	CALMON (DT)	0.8	0.62	0.4	0.08	0.05	0.12
CELIS *	0.83	0.69	0.54	0.18	0.12	0.08	CELIS *	0.83	0.69	0.54	0.18	0.12	0.08
ZAFAR *	0.8	0.65	0.43	0.03	0.18	0.01	ZAFAR *	0.8	0.65	0.43	0.03	0.18	0.01
HARDT (LR)	0.79	0.6	0.53	0.02	0.05	0.06	HARDT (DT)	0.8	0.64	0.5	0.01	0.03	0.07
PLEISS (LR)	0.82	0.73	0.44	0.1	0.09	0.04	PLEISS (DT)	0.82	0.75	0.4	0.08	0.07	0.02
Support Vector Machine (SVM)	0.85	0.75	0.57	0.17	0.073	0.09	Neural Network (NN)	0.85	0.74	0.61	0.19	0.08	0.12
FELDMAN (SVM)	0.81	0.77	0.32	0.06	0.007	0.003	FELDMAN (NN)	0.81	0.7	0.42	0.09	0.015	0.025
CALMON (SVM)	0.78	0.6	0.44	0.03	0.014	0.009	CALMON (NN)	0.78	0.61	0.42	0.08	0.04	0.11
CELIS *	0.83	0.69	0.54	0.18	0.12	0.08	CELIS *	0.83	0.69	0.54	0.18	0.12	0.08
ZAFAR *	0.8	0.65	0.43	0.03	0.18	0.01	ZAFAR *	0.8	0.65	0.43	0.03	0.18	0.01
HARDT (SVM)	0.79	0.63	0.52	0.04	0.01	0.04	HARDT (NN)	0.81	0.63	0.55	0.03	0.003	. 0.06
PLEISS (SVM)	0.82	0.75	0.43	0.1	0.04	0.04	PLEISS (NN)	0.82	0.73	0.4	0.08	0.09	0.03

Method	Accuracy	Precision	Recall	| SPD|	| TPBR |	| FPBR |	Method	Accuracy	Precision	Recall	| SPD|	| TPBR |	| FPBR |
Logistic Regression (LR)	0.65	0.69	0.48	0.28	0.16	0.35	Decision Tree (DT)	0.66	0.66	0.58	0.15	0.089	0.14
FELDMAN (LR)	0.63	0.7	0.39	0.07	0.018	0.069	FELDMAN (DT)	0.64	0.67	0.48	0.034	0.003	0.005
CALMON (LR)	0.66	0.66	0.62	0.12	0.07	0.13	CALMON (DT)	0.68	0.69	0.58	0.06	0.03	0.05
CELIS *	0.66	0.65	0.79	0.18	0.08	0.22	CELIS *	0.66	0.65	0.79	0.18	0.08	0.22
ZAFAR *	0.57	0.55	0.8	0.03	0.02	0.08	ZAFAR *	0.57	0.55	0.8	0.03	0.02	0.08
HARDT (LR)	0.6	0.62	0.39	0.03	0.04	0.09	HARDT (DT)	0.61	0.61	0.48	0.05	0.07	0.03
PLEISS (LR)	0.65	0.69	0.48	0.32	0.41	0.17	PLEISS (DT)	0.65	0.66	0.56	0.22	0.24	0.11
Support Vector Machine (SVM)	0.66	0.67	0.55	0.3	0.2	0.34	Neural Network (NN)	0.67	0.69	0.55	0.32	0.2	0.35
FELDMAN (SVM)	0.64	0.7	0.4	0.08	0.02	0.08	FELDMAN (NN)	0.65	0.69	0.52	0.09	0.009	0.1
CALMON (SVM)	0.67	0.68	0.6	0.18	0.13	0.21	CALMON (NN)	0.67	0.67	0.6	0.14	0.07	0.19
CELIS *	0.66	0.65	0.79	0.18	0.08	0.22	CELIS *	0.66	0.65	0.79	0.18	0.08	0.22
ZAFAR *	0.57	0.55	0.8	0.03	0.02	0.08	ZAFAR *	0.57	0.55	0.8	0.03	0.02	0.08
HARDT (SVM)	0.6	0.62	0.43	0.05	0.08	0.03	HARDT (NN)	0.6	0.62	0.41	0.07	0.06	0.12
PLEISS (SVM)	0.65	0.67	0.54	0.34	0.39	0.21	PLEISS (NN)	0.67	0.71	0.56	0.33	0.33	0.24

Figure 4: (Top): Results for Adult dataset. (Bottom): Results for Compas Dataset. For every machine learning model we report how the various fairness mechanisms affect the behavior of the corresponding classifier in terms of correctness and fairness metrics. $(*)$ Note that the model inside in-processing approaches cannot be changed as fairness constraints within these approaches are strongly coupled with the model. We put the results from in-processing approaches in all the tables for comparison, it does not represent they are using the specific model highlighted in the table (Section 5.2.3 contains more details).

$lr = \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$, and found $lr = 10^{-3}$ to perform well. We used 100 epochs and batch sizes equal to 256 and 512.

5.2.2 Fair Models: Pre-Processing Mechanisms

In our evaluation we use the two pre-processing mechanisms which we described in Section 4: FELDMAN and CALMON. As we mentioned in Section 1, pre-processing mechanisms modify the data to account for fairness. For the implementation of these mechanisms we use the AIF360 toolkit by IBM [3]. We use the publicly available code to generate modified training datasets for the two classification tasks using both CALMON and FELDMAN. On each modified training data, we train four fair classifiers using LR, SVM, DT, and NN as the model. Thus, we have a total of **16** fair approaches across the datasets (as seen in Figure 4).

Figure 2 reports the best hyper-parameter values for the different models we implemented in this project while Figure 3 describes the neural network architectures of the six different models we designed. Finally, Figure 4 reports the correctness and fairness performance of the various methods. We provide a discussion of our results at the end of the subsection.

5.2.3 Fair Models: In-Processing Mechanisms

For implementing the in-processing mechanism of CELIS we use the code in the AIF360 toolkit [3]. For the ZAFAR method we use the publicly available code by the authors. [3] Recall that the in-processing methods are not model-agnostic and typically define a constrained optimization problem that expresses fairness as constraints and are strongly coupled with the model. As a result we used these approaches as is, with no change in the models they come with. ZAFAR builds a logistic regression model with fairness constraints and CELIS builds a customized constrained optimization model using Lagrange duals.

5.2.4 Fair Models: Post-Processing Mechanisms

The post-processing approaches are model-agnostic and modify the predictions in the output of the fairness-unaware model. Consequently, we evaluated the methods of HARDT and PLEISS over all the unfair models we built. For the implementation of the post-processing mechanisms we use the publicly available code from the authors. Specifically, for every unfair model we stored its predictions along with the class probabilities learned for each sample and used the post-processing to modify them. [4]

Correctness and Fairness Evaluation: Discussion of Experimental Findings.

A general trend that we observe in the results in Figure 4 is that the fairness enhancing techniques tend to compromise accuracy over fairness. This finding matches the intuition that since fair approaches divert the primary objective of classification from correctness only to both correctness and fairness. As as a result, some loss in accuracy is expected. However, we also see that in many cases the *price of fairness*, namely the loss in accuracy, is relatively small in most cases.

Further, especially in the case of the Adult dataset we observe a clear trend and see that across all models the different techniques are able to increase the fairness of the machine learning models at least with respect to the fairness metric they are optimizing, if not overall(e.g., FELDMAN and CALMON aim for statistical parity, while HARDT and PLEISS for TPBR and FPBR). We also notice that there is no clear winner over all fairness metrics: e.g., in the DT model FELDMAN has the lowest SPD value but HARDT has the lowest TPBR value. This finding complies with the intuition that since different methods maximize different fairness metrics, we cannot expect them to perform best on metrics for which they were not optimized for. Moreover, the *impossibility of fairness* [7] that states that various fairness notions cannot be simultaneously enforced also supports the results.

Finally, we observe that in the Compas dataset the trend is less clear and some fairness mechanisms even if they minimize some fairness metrics, they often behave worse in some other metric. Consequently, we conclude that although it is possible that some mechanism behaves well in terms of more than one metrics, they are no fairness guarantees on all metrics.

Key takeaway: Fairness usually comes with a penalty in accuracy, as the primary objective of classification is no longer to just maximize prediction accuracy. Further, fair approaches increase fairness w.r.t. to the fairness metric they are designed to optimize, but naturally does not perform best on other metrics. Hence, there is no single best choice of fair approach.

5.3 Efficiency and Scalability Evaluation

In order to evaluate the efficiency and scalability of fair approaches, we did two kinds of experiments based on the Adult dataset as it contained the most number of data points and attributes. The first experiment was to analyze the runtime of fair approaches with increasing amount of data points. We executed a new instance of each approach with a different number of data points (from 1K to 40K) sampled from the dataset. Our second experiment explores the runtime behavior of approaches as the number of attributes increases. We executed a new instance of each approach with a different number of attributes (from 2 to 14). We measure runtime as the time it takes to modify/pre-process the data for pre-processing approaches, the time to train the fair classifier in in-processing approaches, and the time it takes to modify/post-process predictions in post-processing approaches. The reason for

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

[4] https://github.com/gpleiss/equalized_odds_and_calibration

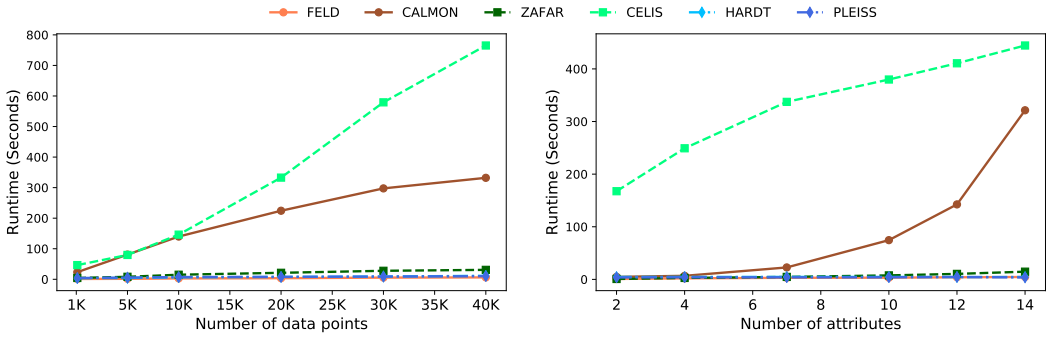

Figure 5: Runtime of the fair approaches with increasing number of data points, and increasing number of attributes.

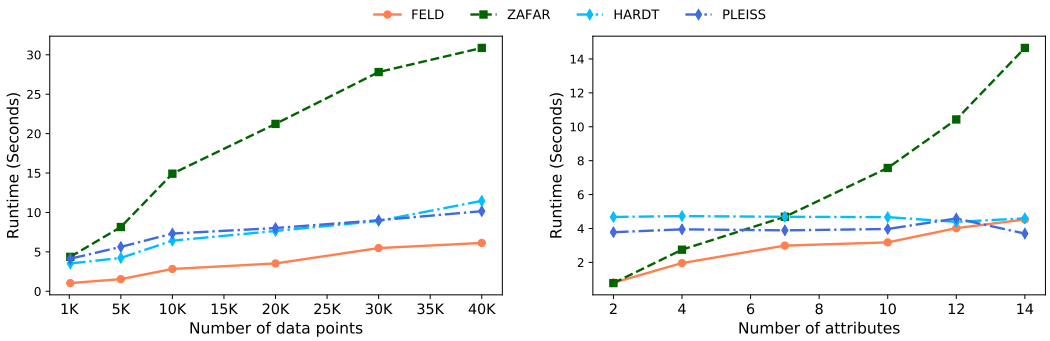

Figure 6: A zoom-in look at the runtime of some of the fair approaches from Figure 5.

not including training time pre- or post-processing is that we wanted to purely report the overhead introduced by these approaches. Our results are shown in Figure 5 and Figure 6.

We notice that different approaches have different approaches scale differently. From figure 5, CALMON and CELIS are the least scalable approaches. CALMON solves a quasi convex optimization problem that has significant runtime requirements. It especially sees almost an exponential increase in runtime as the number of attributes grows. This is due to CALMON being a pre-processing approach that modifies the entire data on per-attribute basis, hence, the complexity of the approach increases significantly with the number of attributes. On the other hand, CELIS is an in-processing approach that also solves a polynomial time constrained optimization problem that also increases in runtime with both the number of data points and the attributes.

Beyond these two approaches, the other approaches are quite efficient, due to their inherent techniques. From a closer look through Figure 6, we notice both of the post-processing approaches are very scalable, this is due to their natural design. Post-processing approaches tend to apply simple modifications to the predictions at the end of the machine learning pipeline and typically never require access to all attributes in training data. ZAFAR is less scalable than the other approaches, although still quite efficient compared CALMON and CELIS.

Key takeaway: The efficiency of pre- and in-processing approaches depend on their core technique, there is no definitive pattern. On the other hand, post-processing approaches are naturally the most scalable and efficient, due to their simplistic mechanisms that only modify the predictions after training.

6 Conclusions and Future Directions

In this work we studied a variety of fair classification mechanisms that have been introduced in the recent literature. We compared their performance in terms of correctness, fairness and scalability across four traditional machine learning models, both linear and non-linear. In terms of correctness and fairness behavior, we observed an overall trend that fairness mechanisms tend to compromise fairness for accuracy, although the loss is relatively small in the general case. Further our experimental results confirm the impossibility of fairness; this notion suggests that different fairness metrics (e.g., SPD, TPBR etc) cannot be satisfied simultaneously. Thus, there is no mechanism that performs best across all metrics. Finally, we observe that post-processing are typically the most scalable since they just depend on the predictions a fairness-unaware model makes and are generally less complex. Consequently, we believe that choosing a fair mechanism highly depends on the task at hand and the fairness metric we are interested in optimizing.

This work focused on a small subset of the available mechanisms that focus on group fairness metrics; for example, we did not consider at all causal based approaches that model the effect the sensitive attribute has on the prediction of the model. As a result a more thorough analysis across mechanisms that also include different fairness definitions could be of independent interest and could serve as a guideline for machine learning scientists.

Updated Collaboration Plan. Maliha Tashfia Islam: Data processing & preparation, experimental design for in-processing and post-processing mechanisms (modified and evaluated CELIS, ZAFAR, PLEISS and HARDT in terms of correctness and fairness metrics), scalability experiments, manuscript writing. Zafeiria Moumoulidou: Data processing & preparation, experimental design for pre-processing mechanisms (hyper-parameter tuning, training, and evaluating the machine models for FELDMAN, CALMON in terms of correctness and fairness metrics), design and implementation of unfair machine learning models (hyper-parameter tuning, training, and evaluating in terms of correctness and fairness metrics), manuscript writing.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[2] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *NIPS Tutorial*, 1, 2017.

[3] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[5] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3995–4004, Red Hook, NY, USA, 2017. Curran Associates Inc.

[6] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

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

[8] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[9] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In

Proceedings of the conference on fairness, accountability, and transparency, pages 329–338, 2019.

[10] S. Galhotra, Y. Brun, and A. Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510, 2017.

[11] R. Goodman. Why Amazon's automated hiring tool discriminated against women. ACLU.org, 2018.

[12] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[13] G. P. Jones, J. M. Hickey, P. G. Di Stefano, C. Dhanjal, L. C. Stoddart, and V. Vasileiou. Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986*, 2020.

[14] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[15] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, 2012.

[16] R. Kohavi and B. Becker. UCI machine learning repository, 1994.

[17] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

[18] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, 2016.

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

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[22] N. Quadrianto and V. Sharmanska. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems*, pages 677–688, 2017.

[23] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[24] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953, 2017.

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

[26] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[27] L. Zhang, Y. Wu, and X. Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3929–3935, 2017.