Can We Obtain Fairness For Free?

Rashidul Islam
University of Maryland, Baltimore County
Baltimore, MD, USA
islam.rashidul@umbc.edu

Shimei Pan
University of Maryland, Baltimore County
Baltimore, MD, USA
shimei@umbc.edu

James R. Foulds
University of Maryland, Baltimore County
Baltimore, MD, USA
jfoulds@umbc.edu

ABSTRACT

There is growing awareness that AI and machine learning systems can in some cases learn to behave in unfair and discriminatory ways with harmful consequences. However, despite an enormous amount of research, techniques for ensuring AI fairness have yet to see widespread deployment in real systems. One of the main barriers is the conventional wisdom that fairness brings a cost in predictive performance metrics such as accuracy which could affect an organization's bottom-line. In this paper we take a closer look at this concern. Clearly fairness/performance trade-offs exist, but are they inevitable? In contrast to the conventional wisdom, we find that it is frequently possible, indeed straightforward, to improve on a trained model's fairness without sacrificing predictive performance. We systematically study the behavior of fair learning algorithms on a range of benchmark datasets, showing that it is possible to improve fairness to some degree with no loss (or even an improvement) in predictive performance via a sensible hyper-parameter selection strategy. Our results reveal a pathway toward increasing the deployment of fair AI methods, with potentially substantial positive real-world impacts.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Machine learning**; **Artificial intelligence**.

KEYWORDS

Fairness in AI, AI & society, deployment of fairness techniques, practical barriers, fairness/performance trade-offs

ACM Reference Format:
Rashidul Islam, Shimei Pan, and James R. Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462614

1 INTRODUCTION

Over the last few years, it has been well established that artificial intelligence (AI) and machine learning (ML) systems, trained on real data without necessary precaution, often exhibit harmful behavior toward certain demographic groups [1, 2, 8, 23, 36]. This can have serious impact on many facets of daily life in a variety of AI-automated tasks including college admissions, resume selections or hiring decisions, financial and housing approvals, criminal justice bail or sentencing decisions, and the prioritization of healthcare resource allocation [37].

With the rising obligation of fairness, the AI community has devoted much effort to the development and enforcement of a broad array of mathematical definitions of fairness in learning algorithms [17, 19–21, 26, 29, 48]. The main paradigm for fair AI/ML models is to posit a quantifiable notion of fairness across protected demographic groups, (e.g. by gender, race, age, etc.) or similar individuals (e.g. persons with similar qualifications and abilities) [4]. The paradigm then enforces these fairness notions by penalizing violations [3, 19, 24] or imposing constraints [47] when optimizing standard machine learning loss functions.

In principle, these techniques should then be able to simply be deployed in the real systems. Unfortunately, the practical reality is far more complicated. Fairness in AI is not a purely technical issue, as it has numerous socio-technical facets crossing computer science/AI, law and policy, the social sciences, and philosophy [6, 22, 41]. Stakeholders who are impacted by these systems are often far-removed or under-represented in the AI research laboratories in academia and industry that design them, yet their voices must be heard [12].

Despite burgeoning research on fairness in AI, learning algorithms which aim to ensure it have currently attained relatively little adoption in deployed AI systems across industry, government, and the public sector. One of the main barriers to the broader adoption of AI fairness in real systems is the potential cost to performance metrics such as accuracy. Since machine learning loss functions and fairness definitions compete to influence a learning algorithm's behavior, there is frequently a trade-off between fairness and predictive performance [11, 33, 50]. A reduction in a deployed model's predictive performance can harm an organization's profitability, which the management may not be willing to tolerate [12]. According to Crawford et al. (2016), "Big Tech refuses to prioritize solving these issues over their bottom line." The expectation of a performance sacrifice also prevents fair AI methods from being considered the default state-of-the-art techniques to be used for any prediction task.

The goal of this work, therefore, is to work toward addressing the "cost of predictive performance" barrier toward deployment of fair AI and ML technologies, in order to ultimately motivate practitioners to increase the adoption of these methods. To that end, we ask a simple question: is it possible to obtain some degree of improvement in fairness metrics *for free*, i.e. without sacrificing performance

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462614

relative to a typical ML model which does not aim to ensure fairness? As we shall see, it turns out that the answer is frequently *yes*. The widely held presumption of a fairness/accuracy trade-off, which is now essentially a "folk theorem" in fair AI research (cf. [4]), neglects the real possibility of achieving a beneficial degree of "fairness for free" in some cases. We identify two mechanisms that can potentially lead to fairness for free: the regularization benefits of fairness penalties, and gerrymandering the errors. We empirically study this "*fairness for free*" phenomenon, and we provide a simple method to achieve it. Our approach is to simply conduct a grid search on the development ("*dev*") set over all model hyper-parameters including the fair ML algorithm's fairness/accuracy trade-off hyper-parameter, in which we optimize for fairness under a performance constraint. Our primary contributions include:

- We systematically study the phenomenon of "fairness for free" using standard fair learning algorithms to resolve the practical limitation of the cost in predictive performance.
- We provide a simple method to simultaneously improve both accuracy and fairness via hyper-parameter tuning.
- Our extensive experimental results on four benchmark datasets demonstrate the benefits of our approach to address one of the major barriers toward deployability of fair AI/ML systems.

2 PRELIMINARIES

In this section, we formalize the general problem setup for fair AI/ML methods, and discuss the algorithms we will use.

2.1 Problem Setup

Fair AI/ML methods typically begin by asserting a mathematical definition which aims to encode a particular notion of fairness; see [4] for an early overview. Much work has been devoted to developing such fairness definitions, including notably demographic parity [17], equalized odds [21], individual fairness [17], counterfactual fairness [29], and intersectional fairness [19, 26]. The paradigmatic AI fairness approach is to formulate training via an objective function $f(\mathrm{X};\theta)$ where a penalty term is added to an ML algorithm's loss function $L(\mathrm{x}_i;\theta)$ which penalizes fairness violations [3]:

$$\min_{\theta} f(\mathrm{X};\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} L(\mathrm{x}_i;\theta) + \lambda F(\mathrm{X};\theta)\,, \tag{1}$$

where N is the number of data points, $\mathrm{x}_i \in \mathrm{X}$ is a data point in the training set, and F is a fairness penalty (typically, lower is better), θ is the model's parameters, and λ is a hyper-parameter that trades between the prediction loss and fairness. In some cases, another approach is used where a constraint on fairness with a small slack tolerance δ is imposed instead of a penalty term [47], which is, roughly speaking, similar to the above up to the choice of λ:

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} L(\mathrm{x}_i;\theta) \qquad \text{s.t. } F(\mathrm{X};\theta) \le \delta\,. \tag{2}$$

There are some methods that do not incorporate fairness measures directly into the objective as a penalty or constraint but rather achieve a fairness criterion implicitly. For example, learning representations of the data that achieve fairness [43, 48] or mitigating unfairness of the ML models by making the predicted output independent from protected attributes such as gender, race, etc. with adversarial training [45, 49]. Although no explicit fairness metric is involved in the learning process, there typically still exists a trade-off parameter λ in these models which balances between fairness and accuracy.

2.2 Fair Learning Algorithms

In this study, we consider two standard fair ML methods, one of which encourages fairness using a penalty term, and the other approach learns fairness implicitly.

2.2.1 Differential Fair Model. The differential fair model (DFM) [19] uses the ϵ-DF differential fairness metric as a penalty term to measure unfairness (lower is better), with regard to parity in the class probabilities assigned to intersecting subgroups of the protected attributes (e.g. Black women). We defer the precise definition of ϵ-DF to Section 6. The learning objective of a deep neural network (DNN)-based classifier $M(\mathrm{x})$ with parameters θ becomes:

$$\min_{\theta} f(\mathrm{X};\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} L(\mathrm{x}_i;\theta) + \lambda[\max(0, \epsilon(\mathrm{X};\theta) - \epsilon_t)]\,, \tag{3}$$

where $\epsilon(\mathrm{X};\theta)$ is the ϵ-DF measures for the classifier and ϵ_t is the desired fairness. If ϵ_t is 0, it penalizes ϵ-DF for $M(\mathrm{x})$. The DFM is trained using stochastic gradient descent [9] on the objective via backpropagation [30] and automatic differentiation [38]. Due to data sparsity of intersectional groups in a minibatch [18], training DFM using stochastic methods is challenging. To address this, a stochastic approximation-based update for $\epsilon(\mathrm{X};\theta)$ is maintained by estimating noisy expected counts per intersecting group for each minibatch [19].

2.2.2 Adversarial Debiasing Model. In the adversarial debiasing model (ADM), an adversarial network penalizes the classifier $M(\mathrm{x})$ if protected attributes z are predictable from the predicted output of the $M(\mathrm{x})$ [31, 45, 49]. In practice, two DNN classification models are used, one with model parameters θ encoding $M(\mathrm{x})$ and an adversary with parameters ϕ, respectively. The learning objective becomes a min-max problem:

$$\min_{\theta}\max_{\phi} f(\mathrm{X};\theta,\phi) \triangleq \frac{1}{N}\sum_{i=1}^{N} L(\mathrm{x}_i;\theta) - \lambda L(\mathrm{X};\theta,\phi)\,, \tag{4}$$

where, the adversary gets the classifier's predictions $\hat{\mathrm{Y}}$ for X instances and attempts to predict z. Both networks are trained simultaneously as follows: the adversary is trained first for an epoch while keeping the classifier fixed, and then the classifier on a minibatch is trained while keeping the adversary fixed [31].

3 FAIRNESS FOR FREE

The conventional wisdom, oft-repeated in AI fairness papers, is that improvements in fairness generally come at a cost in predictive performance [4]. The intuition behind this "folk theorem" is clear from Equation 1, which shows the general formulation that is applied by many AI fairness methods directly or implicitly. In this general objective function, the loss function $L(\mathrm{x};\theta)$ competes with the fairness term $F(\mathrm{X};\theta)$ to determine the desired solution, with an explicit trade-off between the two controlled via a hyper-parameter

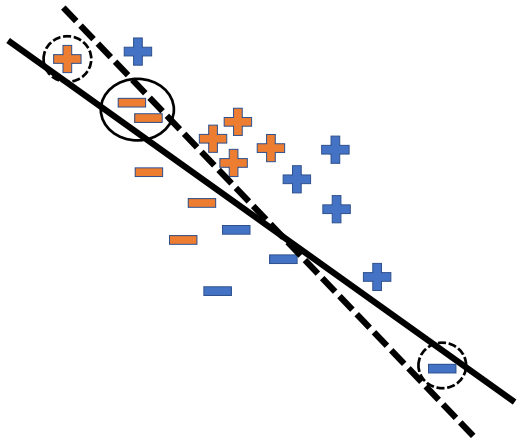

Figure 1: Example of the "fairness for free" phenomenon. Here, + and − are class labels, and colors (orange, blue) correspond to different demographic groups. Solid and dashed circles indicate the errors made by the solid and dashed hyperplanes, respectively.

λ. The higher λ is, the more the fairness penalty is able to shift the optimal solution away from that determined by the predictive loss which we would otherwise solely optimize. This is a serious barrier to the adoption of these methods, because organizations must be persuaded to sacrifice the effectiveness of their systems for the altruistic goal of fair behavior [12].[1]

Although this "folk theorem" generally holds when a strong degree of fairness is required, we will see in this paper that it is often possible to improve fairness to at least some degree with little-to-no loss in accuracy. In our experiments, we demonstrate cases where fair ML models can even improve predictive performance over the equivalent "typical" model (TM). Note that TM is a standard ML model that aims only for accurate predictions, with no fairness intervention.

We will consider two ways in which "fairness for free" can arise. First, Equation 1 has the form of a standard empirical risk minimization objective in which the fairness term is a regularization penalty [3]. Since the fairness penalty acts as a regularizer, it has potential to reduce overfitting, hence improving generalization performance on unseen data while reducing bias to some extent.

Secondly, it is also potentially possible to improve fairness without harming performance on the training data. Multiple different classifiers can potentially obtain the same or a similar number of errors on the training set while making errors on different training instances, a phenomenon known as the Rashomon effect [10]. Some of those equally accurate classifiers may be more fair than others under a desired fairness metric $F(\mathrm{X};\theta)$. We can improve fairness with no loss in performance by selecting the most fair classifier out of these equal-performing models. This can essentially be understood as "gerrymandering" the errors between protected groups at a fixed training-set error rate to optimize fairness which will eventually improve fairness on the unseen data with no overall cost in performance. Figure 1 shows an example where two classifiers have the same number of errors but different fairness. The classifiers with the solid and dashed hyperplanes both make two errors (indicated with solid and dashed circles, respectively). The solid-line classifier makes both errors on orange group instances, while the dashed-line classifier makes one error on an orange group instance and one error on a blue group instance. While this example may seem contrived, practical experience with training deep neural networks, which generally converge to any number of local optima or saddle points of a highly non-convex objective function but obtain similar performance across runs, suggests that for these models there are many different solutions with similar predictive performance [10]. Of these similar quality solutions, we can and should aim to pick the most fair option.

So, the "fairness for free" phenomenon can potentially be achieved with fair methods that improve fairness with no loss in predictive performance by 1) reducing overfitting, and/or 2) "gerrymandering" the training errors between protected groups. We consider two strategies to obtain "fairness for free" models by simply using standard fair ML methods. In both methods, we first obtain a *typical model (TM)* via a grid search over TM's hyper-parameters (e.g. # neurons/layer, activation function, drop-out probability, etc.), selecting the best TM based solely on a performance metric such as accuracy, measured on the development set. Our goal is to find a model which improves fairness over TM, while retaining or improving its predictive performance.

3.1 Full Hyper-parameter Search (FHS)

In the Full Hyper-parameter Search (FHS) method, which we consider to be our gold-standard approach, we select the best fair models in terms of performance and fairness metrics on the development set, via a *grid search over all hyper-parameters*, including those for TM, *and* the trade-off hyper-parameter λ. We select the model with the best fairness metric, such that the performance metric is at least as good as for TM. Note that a feasible model satisfying the constraint always exists, since $\lambda = 0$ corresponds to TM.

3.2 Stage-wise Hyper-parameter Search (SHS)

As a faster alternative to FHS, in the Stage-wise Hyper-parameter Search (SHS) method the fair model is assigned the same hyper-parameter values as the best TM. Then, a grid search is conducted over *only* the fairness trade-off λ, holding the other hyper-parameters fixed. We select the model with the best fairness metric, such that the performance metric is at least as good as for TM.

4 EXPERIMENTS

We conduct an extensive experimental analysis to study whether fair learning algorithms can enforce fairness while retaining or improving predictive performance. The implementation's source

[1] There are good non-altruistic reasons to adopt fair AI, such as avoiding legal liability regarding anti-discrimination laws such as Title VII of the Civil Rights Act of 1964, or to improve the organization's reputation as an ethical actor. As of now, these reasons have not been sufficient for widespread adoption [12].

code of our study is provided on GitHub.[2] All experiments were performed on the following benchmark datasets:

- **COMPAS:** The COMPAS dataset regarding a system which is used to predict criminal recidivism. It has been criticized as potentially biased [1]. Following [18], we used *race* (4 values: *black*, *white*, *hispanic*, and *others*) and *gender* (binary: *men* and *women*) as protected attributes. The target variable indicates "actual recidivism," which is binary, within a 2-year period for 7.22K individuals.
- **Adult:** The Adult 1994 U.S. census income dataset from the UCI ML-repository [15] consists of 14 attributes such as work, relationships, and demographics for individuals, pre-split into a training set of 32.56K instances and a test set of 16.28K instances. The downstream task is to predict whether an individual earns more than $50K/year. Following [18], we selected *race* (4 values: *black*, *white*, *asian-pac-islander*, and *others*), *gender* (binary: *men* and *women*), and *nationality* (binary: *U.S.* and *others*) as the protected attributes.
- **Bank:** The bank marketing data, extracted from direct marketing campaigns of a Portuguese bank [34], contains a total of 41.18K subjects, and each with 20 attributes. In addition to *age* (binary: $age < 35$ and $age \geq 35$) by following [46], we used *job* (binary: *privileged* and *unprivileged*) as protected attributes. The downstream task is to predict whether the client has subscribed or not to a term deposit.
- **HHP:** A dataset derived from the Heritage Health Prize (HHP) milestone 1 challenge,[3] a considerably larger dataset which contains information for 171.07K patients over a 3 year period. The task is to predict whether the *Charlson Index*, an estimation of mortality, is greater than zero. Following [43], we also used *age* (9 values: $55 \leq age < 65$, $65 \leq age < 75$, etc.) and *gender* (binary: *men* and *women*) as the protected attributes.

4.1 Experimental Settings

We investigate and compare the FHS and SHS strategies using two standard fair models, DFM and ADM, with the TM baseline (no fairness intervention). All the models were trained via adaptive gradient descent optimization (Adam) [28] using PyTorch [39] for a total of 10 epochs. Table 1 summarizes the set of hyper-parameter[4] values to perform the grid search on the development (dev) set. Note that the DNN-based hyper-parameters (all hyper-parameters except λ) are common for all models. Further note that, we used same network configurations for classifier and adversarial networks in ADM, following [31]. Since DFM requires a relatively smaller λ in practice, we used different ranges for λ values in DFM and ADM, respectively, as shown in Table 1. It is also intelligible from the table that we trained 96 TM models with every grid for DNN-based hyper-parameters to pick the best option. For fair models (both DFM and ADM), we trained 10 and 960 models following SHS (requires tuning for λ only) and FHS (requires tuning for DNN-based hyper-parameters plus λ) strategies, respectively.

[2]https://github.com/rashid-islam/F3_via_grid.
[3]https://www.kaggle.com/c/hhp.
[4]We refer to [30] for details on the deep learning networks and effect of hyper-parameters on them.

#neurons / hidden layer	{[64, 64, 64], [32, 32, 32], [64, 32, 16]}
minibatch size	{128, 256}
learning rate	{0.001, 0.005}
dropout probability	{0, 0.5}
activation function	{ReLU, LeakyReLU}
$l2$ regularization	{0, $1e$-5}
λ for DFM	10 evenly spaced values in [$1e$-5, 0.1]
λ for ADM	10 evenly spaced values in [1, 10]

Table 1: Set of hyper-parameter values for the grid search.

We split the COMPAS and Bank datasets into 60% train, 20% dev, and 20% test sets. For the Adult data, we used the pre-specified test set and held-out 30% from the training data as the dev set. Finally, we held-out 10% from our larger data HHP as the test set, using the remainder for training, while 10% from the training set was further held-out as the dev set.

To evaluate the predictive performance of the models, we computed accuracy, F1 score, and ROC AUC [25] for the held-out data. For fairness measures, we computed ϵ-DF and γ-SF with all protected attributes, while δ-DP and $p\%$-Rule [47] were measured for each protected attribute, e.g. *gender*, *race*, *age*, etc., separately (see section 6 for details). Since, by definition, δ-DP and $p\%$-Rule assume binary protected attributes between privileged and unprivileged groups, we converted non-binary protected attributes into binary for these two metrics. For example, *race* is coded as *white* and *non-white* for COMPAS and Adult, while *age* are coded as $age \leq 65$ and $age > 65$ for HHP data. Furthermore, we select the most marginalized subgroup (*mmsg*) as the protected group (*black women non-USA* for Adult, *black women* for COMPAS, $age < 35$ with *unprivileged job* for Bank, and *women* with $age \geq 85$ for HHP) for δ-DP and $p\%$-Rule measurements, compared to its complement.

4.2 Analysis on Grid Search

Since AI fairness interventions divert a system's learning objective from accuracy only to both accuracy and fairness, the conventional wisdom is it may hurt accuracy. In this experiment, we study the impact of fairness interventions on the accuracy with respect to the hyper-parameters for all benchmark datasets following our SHS and FHS approaches.

Figure 2 shows accuracy versus various fairness metrics for all models on the dev set of COMPAS. The best option for the TM baseline, selected in terms of highest accuracy on the dev set via grid search over hyper-parameters, is indicated by a *black asterisk*. All fair models obtained during the grid search are shown, including DFM (*blue circles*) and ADM (*purple diamonds*). The "fairness for free" region (*orange area*) is marked by an area that has equal or higher accuracy and better corresponding fairness than the best TM. Depending on the fairness metric, lower (left) or higher (right) may be better. **With the FHS approach, a large number of fair models, both DFM and ADM, satisfied the criteria of "fairness for free"** in terms of all of the fairness metrics.

The SHS approach did not perform as well in this experiment. We found only a single fair ADM model which satisfied our criteria of "fairness for free" for all fairness metrics, except δ-DP (*mmsg*) and $p\%$-Rule (*mmsg*).

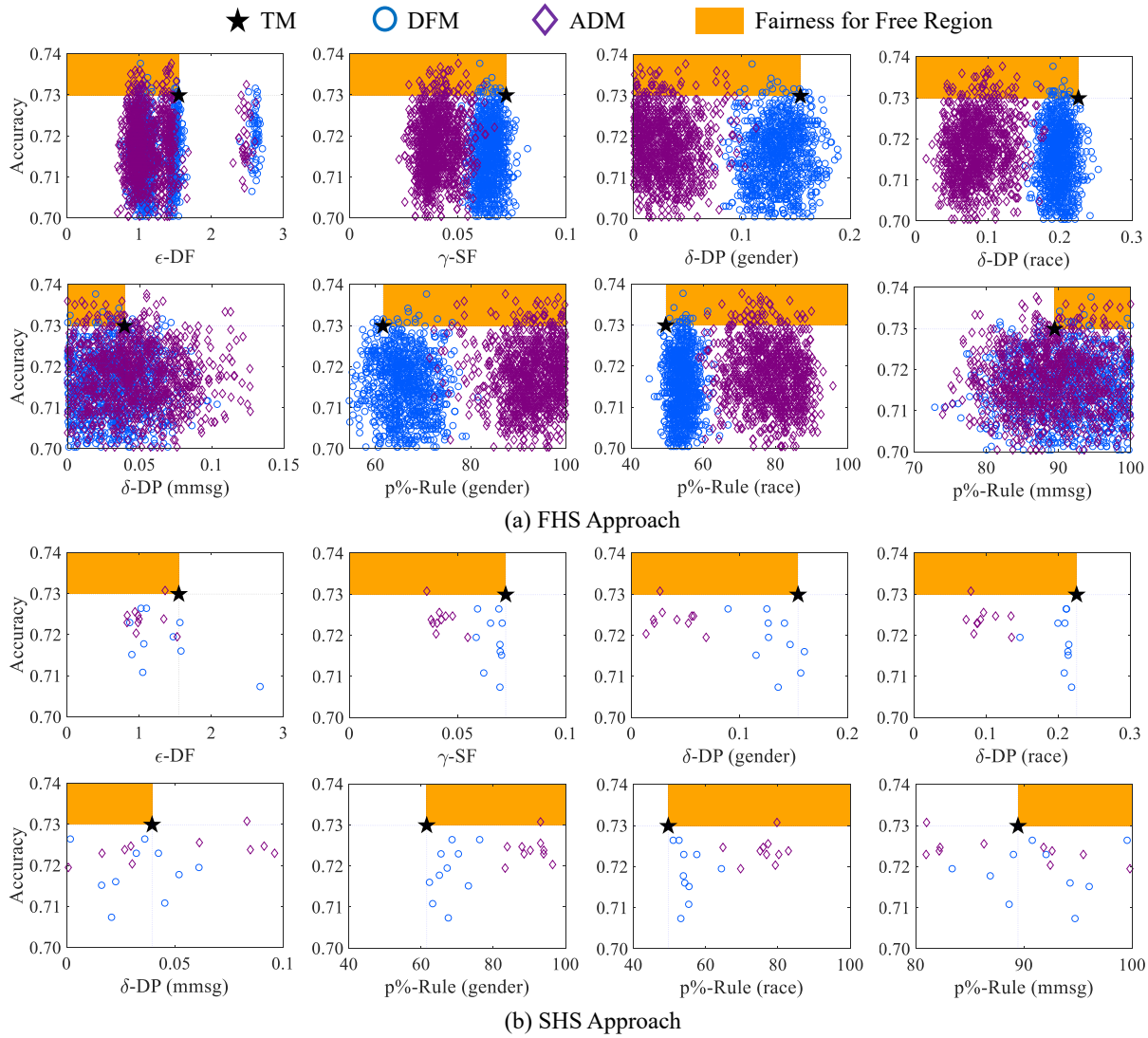

Figure 2: Analysis on the grid search for the dev set of COMPAS dataset. *Black asterisk* is the best typical model (TM) in terms of accuracy, while the *blue circle* and *purple diamond* represent differential fair models (DFM) and adversarial debiasing models (ADM), respectively, trained on hyper-parameter grids with (a) full hyper-parameter search (FHS) and (b) stage-wise hyper-parameter search (SHS). A large number of fair models satisfy the "fairness for free" (*orange area*) criteria using the FHS approach with respect to all fairness metrics, while SHS only satisfied the criteria in a few cases. Lower is better for ϵ-DF, γ-SF, and δ-DP; higher is better for p%-Rule.

Similar conclusions can be found on the Adult and Bank datasets (Figures 3 and 4). For these datasets, FHS again achieved a considerable number of "fairness for free" cases, although fewer than for COMPAS, while the *orange area* was empty for SHS (i.e. fairness for free was not achieved by the SHS method in this case).

As shown in Figure 2 and 3, there are relatively more FHS fair models in the "fairness for free" region in case of the COMPAS dataset compared to Adult. As a likely explanation, we observe a higher difference in accuracy between train and dev sets for the TM baseline on this dataset, which created more scope for fair models to improve both accuracy and fairness via regularization. For example, to improve fairness, fair models may alter the TM baseline's predicted false negative individuals of an unprivileged groups to true positives which provide improvement in accuracy as well. To further study the generalization behavior of the models, we provide a case study for COMPAS dataset in a later section.

4.3 Performance of "Fairness for Free" Methods

We next evaluated the accuracy-based performance and fairness metrics for the best fair models, with respect to "fairness for free" phenomenon, on *unseen test data*, and compared them with the TM baseline. In FHS approach, we picked the fairest model on the dev set, under the corresponding fairness metric, that provides equal or

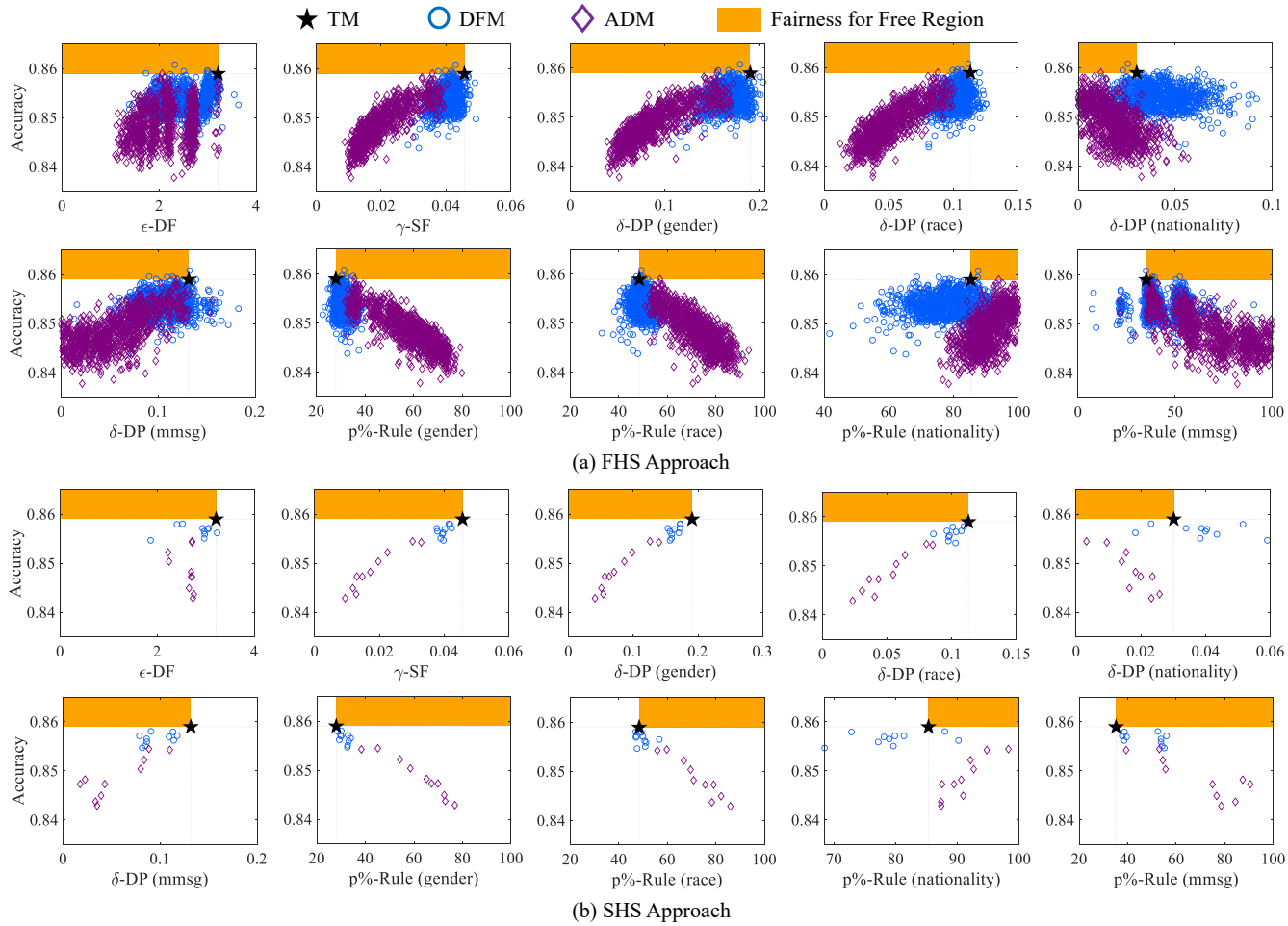

Figure 3: Analysis on the grid search for the dev set of Adult dataset. *Black asterisk* is the best typical model (TM) in terms of accuracy, while *blue circles* and *purple diamonds* represent differential fair models (DFM) and adversarial debiasing models (ADM), respectively, trained on hyper-parameter grids with (a) full hyper-parameter search (FHS) and (b) stage-wise hyper-parameter search (SHS). A considerable number of fair models satisfy the "fairness for free" (*orange area*) criterion using the FHS approach with respect to all fairness metrics, while no fair model satisfies the criteria using the SHS approach. Lower is better for ϵ-DF, γ-SF, and δ-DP; higher is better for p%-Rule.

higher accuracy compared to TM. Since very few SHS fair models reached the "fairness for free" region on the dev set (see Figure 2, 3, and 4), we relaxed the accuracy constraint. For SHS, we instead selected the best DFM and ADM models in terms of their accuracy on the dev set that also ensures improved fairness over TM, under the corresponding fairness metric. As DFM and ADM are optimized to ensure ϵ-DF and δ-DP (or p%-Rule), respectively, in its learning process, we treated ϵ-DF and δ-DP (*mmsg*) as their corresponding fairness metrics, respectively.

In Table 2, we show detailed results for the selected TM baseline, fair models using SHS (DFM-S and ADM-S), and fair models using FHS (DFM-F and ADM-F). In all datasets, DFM-F and ADM-F were the best models overall in terms of accuracy, while both of them improved all the fairness metrics comparing to TM. Furthermore, ADM-F performed with the highest accuracy, outperformed all other models on all datasets, and ensured better fairness than TM as well. Since the accuracy constraint was relaxed in the model selection criteria on the dev set for the SHS approach, DFM-S and ADM-S improved fairness metrics to a higher degree for most of the cases, with little-to-no loss in accuracy compared to TM.

For the COMPAS dataset, ADM-F was the fairest model in terms of δ-DP (*race*), p%-Rule (*gender*), and p%-Rule (*race*) in addition to performing with highest accuracy, while ADM-S was the fairest model in term of all other fairness measures with similar performance in accuracy to TM. Though DFM-S outperformed the others with respect to F1 score, it increased unfairness compared to TM in terms of ϵ-DF, γ-SF, and δ-DP (*gender*). Worsening some fairness metrics with DFM-S was unexpected since the DFM-S showed better fairness on the dev set. This counter-intuitive result is presumably due to the difference in the distribution of protected groups for the train and dev sets. In the case of the Adult dataset, ADM-F achieves the highest improvement in accuracy and ϵ-DF, while ADM-S was the fairest model in terms of all the other fairness metrics with a little loss in accuracy. DFM-F showed superior performance on the

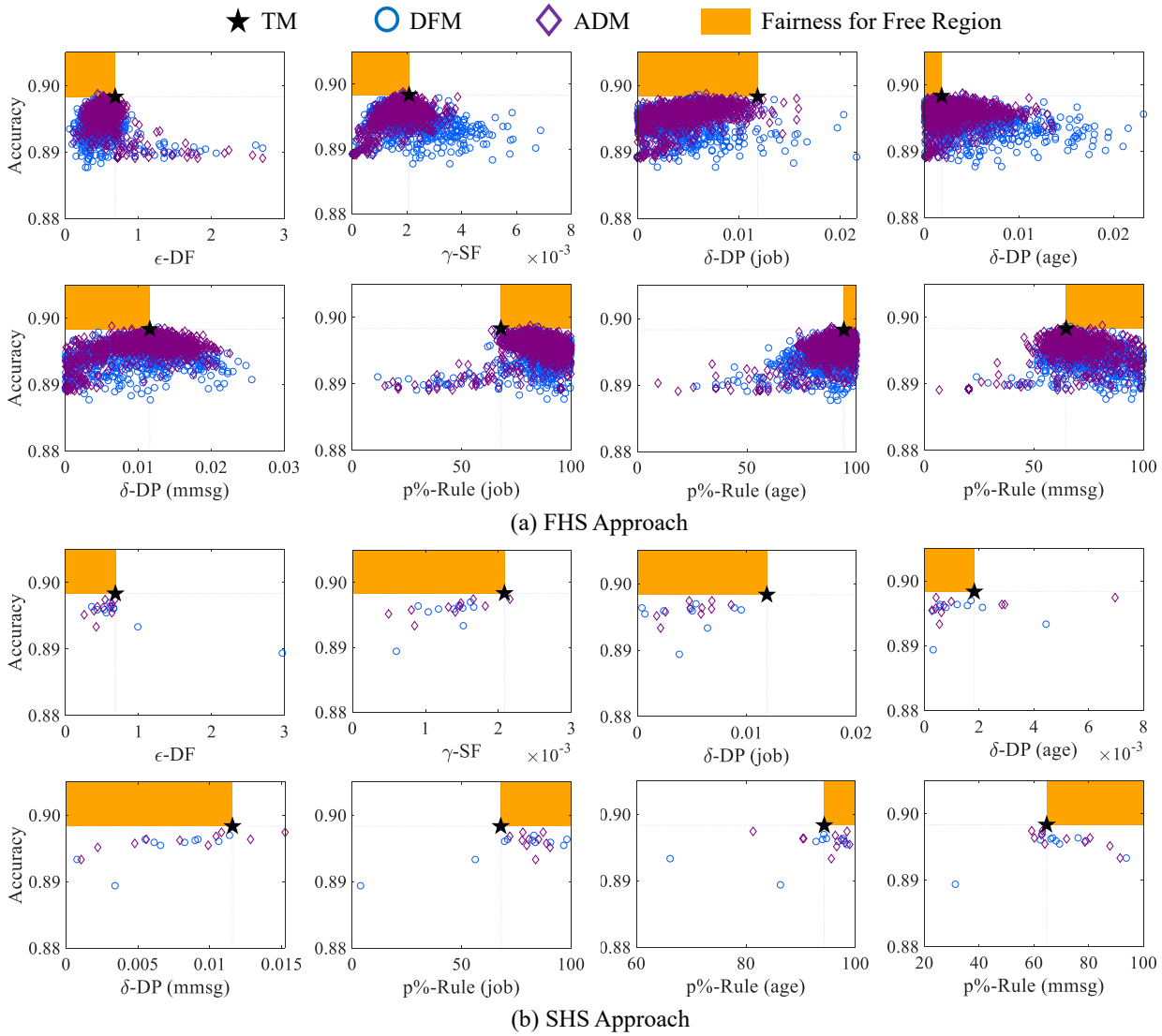

Figure 4: Analysis on the grid search for the dev set of Bank dataset. *Black asterisk* is the best typical model (TM) in terms of accuracy, while *blue circle* and *purple diamond* represent differential fair models (DFM) and adversarial debiasing models (ADM), respectively, trained on hyper-parameter grids with (a) full hyper-parameter search (FHS) and (b) stage-wise hyper-parameter search (SHS). A considerable number of fair models satisfy "fairness for free" (*orange area*) criteria using FHS approach with respect to all fairness metrics, while no fair model satisfy the criteria using SHS approach. Lower is better for ϵ-DF, γ-SF, and δ-DP; higher is better for p%-Rule.

Bank dataset in terms of all predictive performance metrics and most of the fairness metrics. Finally, on the HHP dataset, we get the highest improvement in accuracy and F1 score using ADM-F, with a considerable amount of improvement in all fairness metrics, while ADM-S shows huge improvement in most of the fairness metrics with a loss in predictive performance.

4.4 Case Study on Overfitting for COMPAS

The COMPAS system predicts recidivism score to determine sentencing and supervision for incarcerated individuals, and which has been criticized as potentially racially biased [1]. Compared to the other datasets in our experiments, we found that ML models often exhibit more overfitting on COMPAS. This presents an opportunity for fair models to improve both accuracy and fairness using our approach.

To investigate this further, in Figure 5 we compared the generalization of the TM with fair models DFM and ADM on the train and test sets of COMPAS in terms of accuracy and p%-Rule (*race*) while varying model complexity. In this experiment, we varied the number of hidden layers of all models, where each hidden layer contained a fixed number of neurons, e.g. 64. To demonstrate the impact of fairness interventions on the overfitting, we removed hyper-parameters that compensate for overfitting such as dropout probability and $l2$ regularization by setting them to 0. At every

COMPAS Dataset											
Models	Accuracy ↑	F1 Score ↑	AUC ↑	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (gender)	δ-DP ↓ (race)	δ-DP ↓ (mmsg)	p%-Rule ↑ (gender)	p%-Rule ↑ (race)	p%-Rule ↑ (mmsg)
TM	0.688	0.635	**0.729**	1.793	0.071	0.234	0.145	0.177	46.772	67.372	56.812
DFM-S	0.691	**0.644**	0.727	1.824	0.072	0.244	0.140	0.176	46.578	69.376	58.369
ADM-S	0.688	0.616	0.721	**1.580**	**0.043**	**0.138**	0.045	**0.123**	63.612	87.863	**66.170**
DFM-F	0.688	0.621	**0.729**	1.659	0.054	0.203	0.102	0.150	49.395	74.177	59.960
ADM-F	**0.694**	0.633	0.726	1.632	0.044	0.143	**0.026**	0.156	**64.387**	**93.295**	59.928

Adult Dataset													
Models	Accuracy ↑	F1 Score ↑	AUC ↑	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (gender)	δ-DP ↓ (race)	δ-DP ↓ (nationality)	δ-DP ↓ (mmsg)	p%-Rule ↑ (gender)	p%-Rule ↑ (race)	p%-Rule ↑ (nationality)	p%-Rule ↑ (mmsg)
TM	0.854	**0.665**	**0.907**	2.679	0.046	0.184	0.101	0.060	0.150	29.194	52.506	70.739	24.615
DFM-S	0.854	0.655	**0.907**	2.501	0.042	0.167	0.098	0.061	0.138	30.950	51.243	68.266	26.282
ADM-S	0.852	0.647	0.899	1.692	**0.030**	**0.120**	**0.071**	**0.030**	**0.115**	**46.315**	**63.516**	**83.848**	**37.300**
DFM-F	0.854	0.662	**0.907**	2.564	0.042	0.168	0.091	0.065	0.128	33.538	56.526	68.033	34.882
ADM-F	**0.856**	0.661	0.906	**1.850**	0.036	0.149	0.076	0.041	0.121	37.571	61.837	78.908	36.279

Bank Dataset											
Models	Accuracy ↑	F1 Score ↑	AUC ↑	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (age)	δ-DP ↓ (job)	δ-DP ↓ (mmsg)	p%-Rule ↑ (age)	p%-Rule ↑ (job)	p%-Rule ↑ (mmsg)
TM	**0.897**	0.288	0.754	0.802	0.004	0.010	0.017	0.004	73.864	57.819	88.739
DFM-S	**0.897**	0.273	0.760	0.436	**0.002**	**0.003**	0.011	0.005	**92.245**	68.199	82.542
ADM-S	**0.897**	0.278	0.757	0.594	0.003	0.005	0.014	0.006	86.221	61.647	79.007
DFM-F	**0.897**	**0.312**	**0.761**	**0.431**	**0.002**	0.008	**0.009**	0.001	80.289	**78.730**	97.727
ADM-F	**0.897**	0.284	0.760	0.563	0.003	0.008	0.011	**0.000**	78.463	69.908	**98.468**

HHP Dataset											
Models	Accuracy ↑	F1 Score ↑	AUC ↑	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (gender)	δ-DP ↓ (age)	δ-DP ↓ (mmsg)	p%-Rule ↑ (gender)	p%-Rule ↑ (age)	p%-Rule ↑ (mmsg)
TM	0.860	0.752	0.916	2.746	0.026	0.023	0.397	0.280	91.382	21.654	40.997
DFM-S	0.862	0.755	**0.917**	2.720	0.025	0.020	0.398	0.285	92.419	21.546	40.433
ADM-S	0.855	0.741	0.902	**2.512**	**0.021**	0.024	**0.308**	**0.216**	91.096	**31.273**	**48.841**
DFM-F	0.863	0.756	0.915	2.640	0.024	**0.015**	0.382	0.272	**94.107**	23.079	41.848
ADM-F	**0.864**	**0.760**	0.915	2.684	0.024	0.023	0.389	0.269	91.337	22.965	42.708

Table 2: Performance for best fair models in terms of "fairness for free" phenomenon on the test set of all datasets. Fair models using full hyper-parameter search (FHS) performed with highest accuracy, and also improved all fairness metrics to some degree. Fair models using stage-wise hyper-parameter search (SHS) provided higher improvement in fairness metrics for most of the cases with little-to-no loss in accuracy. TM: typical model; DFM-S and DFM-F: differential fair model using SHS and FHS, respectively; ADM-S and ADM-F: adversarial debiasing model using SHS and FHS, respectively. Higher is better for measures with ↑, while lower is better for measures with ↓.

configuration of hidden layers, the best options for the models were selected again via grid search on the dev set over the rest of the hyper-parameters using our FHS approach. Figure 5 shows that the accuracy of TM is higher on the train set comparing to fair models when network size increases, while both fair models outperform TM on the test set with any network size. Furthermore, DFM and ADM decrease the corresponding gap between accuracy on train and test sets due to the regularization behavior of the fairness interventions. As also shown, DFM and ADM ensure higher p%-Rule (*race*) on both train and test sets compared to TM. Since fairness interventions affect the learning objective in the training phase, both DFM and ADM exhibit fairness improvement on the train set as well. Overall, we conclude from this experiment that **fair models reduce overfitting which helps to improve both accuracy and fairness.**

5 DISCUSSION

In this work, we investigated how some degree of "fairness for free," where fairness improvements do not harm (and perhaps improve) prediction, can be achieved by standard fair learning algorithms. We provided two strategies, SHS and FHS, for conducting a grid search over hyper-parameters to find the "fairness for free" models. The methods are applicable to any fair models that use a trade-off

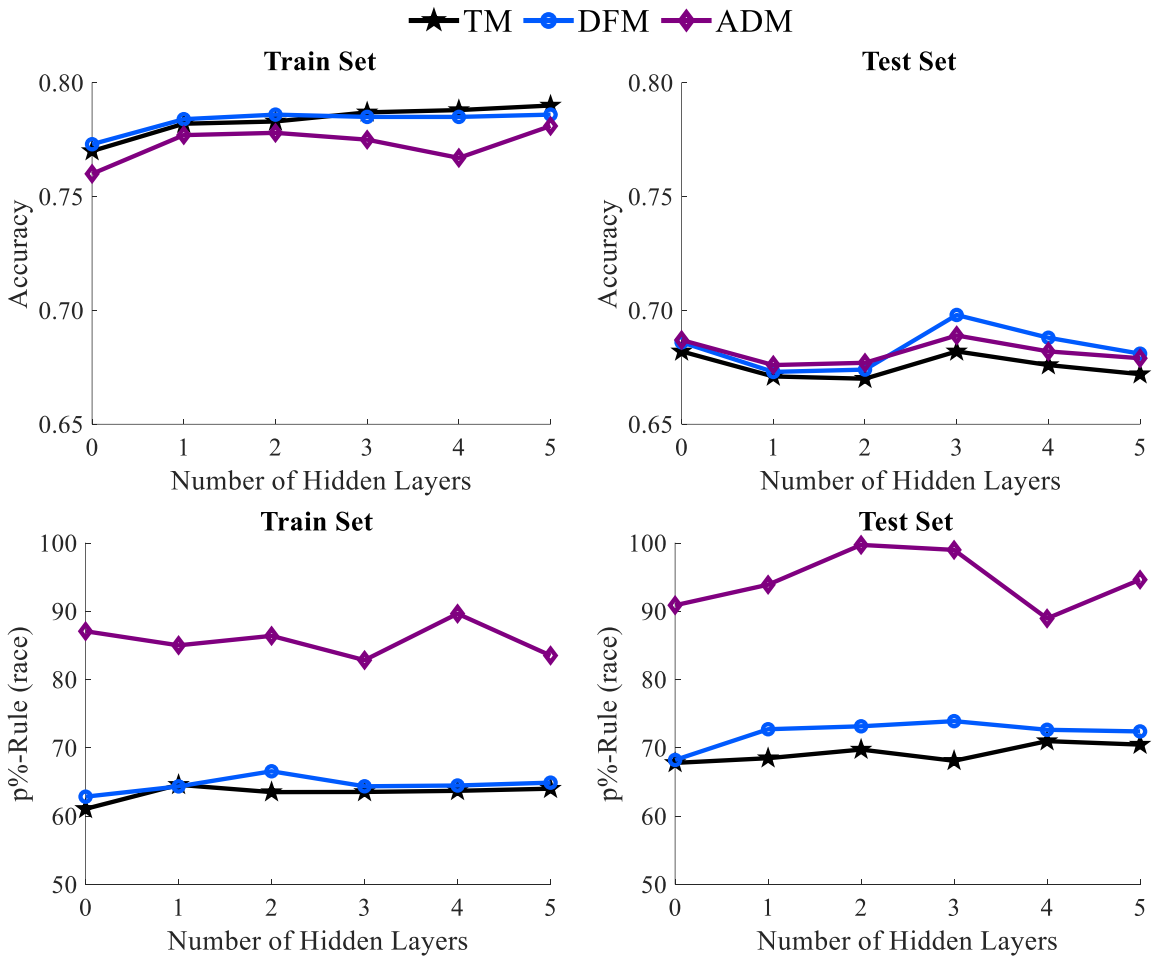

Figure 5: Comparison of the generalization for the typical model (TM), differential fair model (DFM), and adversarial debiasing model (ADM) in terms of accuracy and p%-Rule (*race*) on the train and test set of COMPAS dataset, while varying the network size. Higher is better for accuracy and p%-Rule (*race*). The results indicate that DFM and ADM reduced overfitting compared to TM.

parameter in the learning process to balance between the predictive performance and fairness. As shown in Table 2, **fair models using FHS outperformed TM in terms of accuracy, and improved all fairness measures considered** including group fairness, δ-DP and p%-Rule, and intersectional fairness metrics, e.g. ϵ-DF and γ-SF. Our experiments on multiple benchmark datasets challenge the traditional perception of the accuracy and fairness trade-off.

The main limitations of our FHS approach is that it is computationally expensive since the fair models must be trained many times to pick the the best one. Our other approach **SHS is relatively less computationally expensive and substantially improves most of the fairness metrics with little-to-no loss in predictive performance**. However, we had a few failure cases with SHS approach, e.g. DFM using SHS slightly increased unfairness in terms of ϵ-DF, γ-SF, δ-DP (*gender*), p%-Rule (*gender*) on COMPAS, and δ-DP (*nationality*), p%-Rule (*nationality*) on Adult (see Table 2). Furthermore, our approaches are only applicable to models for which fair learning algorithms have already been developed.

In future, we plan to address these limitations, e.g. speeding up our hyper-parameter tuning approach using Bayesian optimization [42] to carefully and automatically tune the hyper-parameters. We also aim to develop a learning algorithm that directly solves a constrained optimization problem to obtain "fairness for free." Furthermore, we will conduct user studies on our developed methods to investigate how the individuals from multiple disciplines perceive and interact with our developed methods. The eventual goal of this research is the successful application and deployment of our methods, while preserving public trust in AI systems.

6 RELATED WORK

In our approach, we optimized fairness under an accuracy constraint. This reverses a common approach in which predictive performance is optimized under a fairness constraint, exemplified by the work of [47]. For example, Perrone et al. 2020 proposes to use Bayesian optimization to select the hyper-parameters of any black-box ML model to optimize accuracy within a fairness constraint.

The existence of the fairness/accuracy trade-off has been identified and characterized in several existing works [11, 33, 50]. Ustun et al. 2019 aim to select ML models specific to each demographic group, e.g. a group-level model vs an overall "pooled model" of all groups, such that the models perform as accurately as possible for that group, without harming that group. In [19], the fairness/accuracy trade-off hyper-parameter is selected for fair models based on the validation set to minimize unfairness, such that accuracy is allowed to be degraded by at most 5% from the typical model.

The recent work of [16] questions the legitimacy of the fairness / accuracy trade-off. They argue that disparities in predicted outcomes may arise from historical and systemic biases, and that predictive performance on such biased data is less desirable than performance on an idealized debiased data distribution. They then devise an optimization method to obtain such a distribution. There are several other studies [7, 32] that challenge the fairness/accuracy trade-off by decomposing the bias in the training data into two parts (e.g. recoverable and non-recoverable), and demonstrate that enforcing fairness criteria improve performance of the Bayes optimal model for recoverable data part. Our approach differs in that we intervene on the existing fair learning algorithms rather than on the data distribution, and that we aim to prevent performance degredation on the original data rather than on any modified data. This work was motivated by our recent findings regarding the prevention of overfitting in survival analysis with fair survival models for equitable allocation of medical resources [27, 35].

6.1 Fairness Metrics

In this section we provide definitions for the fairness metrics used in our experiments. We selected standard fairness metrics for protecting groups, e.g. *men*, *women*, *black*, *white*, as well as those which protect intersectional groups, e.g. *black women* and *white men*. We assume a finite dataset of N individuals in which each individual is defined as a triple of attributes x, corresponding ground-truth class y, and protected attributes z which might be included in x. Let $M(\mathrm{x})$ be an algorithmic mechanism (e.g. classifier) which takes an instance x and assigns them an outcome $\hat{y}$, e.g. whether or not the individual was awarded a loan.

6.1.1 Demographic Parity. The demographic parity (DP) [13, 17, 48] criterion is satisfied when the predictions $\hat{y}$ are independent of the protected attribute z, where z is assumed to be a binary variable. Since it is often impossible to achieve complete independence, a practical metric, a demographic parity distance δ [13], is defined. Specifically:

A mechanism $M(\mathrm{x})$ satisfies δ-DP with respect to $z \in \{0,1\}$ if

$$|P(M(\mathrm{x})=1|z=1) - P(M(\mathrm{x})=1|z=0)| \leq \delta\,, \tag{5}$$

where $z=1$ and $z=0$ indicates privileged and unprivileged groups, respectively. Smaller δ is better, e.g. $\delta = 0$ indicates absolute fairness.

Alternatively, by considering the ratio between groups' outcome probabilities instead of their absolute difference, we obtain the $p\%$-Rule [47], which generalizes the 80% rule of the U.S. employment law [5], that measures disparate impact toward a protected group. *A mechanism $M(\mathrm{x})$ satisfies the $p\%$-Rule if*

$$\min\left(\frac{P(M(\mathrm{x})=1|z=1)}{P(M(\mathrm{x})=1|z=0)}, \frac{P(M(\mathrm{x})=1|z=0)}{P(M(\mathrm{x})=1|z=1)}\right) \geq \frac{p}{100}\,, \tag{6}$$

where larger p is better, e.g. $p = 100\%$ represents perfect fairness, otherwise $p < 100\%$.

6.1.2 Subgroup Fairness. The subgroup fairness (SF) metric [26] is a multi-attribute definition with respect to all intersectional subgroups (e.g. *black women*) and top-level groups (e.g. men). *Let $\mathcal{G}$ be a collection of protected group indicators $g : A \rightarrow \{0,1\}$, where $g(\mathrm{s}) = 1$ designates that an individual with protected attributes s is in group g. Then $M(\mathrm{x})$ satisfies γ-SF with respect to $\mathcal{G}$ if for every $g \in \mathcal{G}$, and $\hat{y} \in \{0,1\}$,*

$$|P(M(\mathrm{x})=1) - P(M(\mathrm{x})=1|g(\mathrm{s})=1)| \times P(g(\mathrm{s})=1) \leq \gamma\,, \tag{7}$$

where $\gamma \in [0,1]$, and smaller is better. Since the term $P(g(\mathrm{s})=1)$ weights the penalty by the size of group g as a proportion of the population, γ-SF does not guarantee to protect the small minority groups [19].

6.1.3 Differential Fairness. The differential fairness (DF) metric [19] is a definition specifically motivated by intersectionality [14], which aims to ensure equitable treatment by an algorithm for all intersecting subgroups of a set of protected categories with additional beneficial properties from a societal perspective regarding the law, privacy, and economics.

Let $s_1, \ldots, s_p$ be discrete-valued protected attributes, $z = s_1 \times s_2 \times \ldots \times s_p$. A mechanism $M(\mathrm{x})$ satisfies ϵ-DF with respect to z if for all x, and $\hat{y} \in Range(M)$,

$$e^{-\epsilon} \leq \frac{P(M(\mathrm{x})=\hat{y}|\mathrm{s}_i)}{P(M(\mathrm{x})=\hat{y}|\mathrm{s}_j} \leq e^{\epsilon}\,, \tag{8}$$

for all $(\mathrm{s}_i, \mathrm{s}_j) \in z \times z$ where $P(\mathrm{s}_i) > 0, P(\mathrm{s}_j) > 0$. Smaller ϵ is better, and $\epsilon = 0$ for perfect fairness, otherwise $\epsilon > 0$.

7 CONCLUSION

We investigated one of the major practical barriers to the wide deployment of fair AI/ML systems across industry, government, and the public sector: the cost in predictive performance when ensuring fairness. We showed experimentally that it is frequently possible to improve fairness to some degree while avoiding any reduction in predictive performance. This can be achieved with any standard fair learning algorithm by using a simple but sensible approach to hyperparameter tuning. Our extensive experimental results on multiple benchmark datasets demonstrate the practicality of the proposed techniques. We hope that our results will encourage further research in addressing the human-facing barriers to deployment of fair AI methods, leading to increased real-world societal benefit from these technologies.

ACKNOWLEDGMENTS

This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology. This material is based upon work supported by the National Science Foundation under Grant No.'s IIS2046381; IIS1850023; IIS1927486. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* 23 (2016).

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

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *4th Annual Workshop on Fairness, Accountability, and Transparency in Machine Learning. ArXiv preprint arXiv:1706.02409 [cs.LG]* (2017).

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

[5] Dan Biddle. 2006. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing.* Gower Publishing, Ltd.

[6] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency.* PMLR, 149–159.

[7] Avrim Blum and Kevin Stangl. 2020. Recovering from biased data: Can fairness constraints improve accuracy?. In *1st Symposium on Foundations of Responsible Computing.* Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems.*

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

[10] Leo Breiman et al. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[11] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems.* 3539–3550.

[12] Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times* (2016).

[13] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning.* PMLR, 1436–1445.

[14] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.* (1989), 139.

[15] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[16] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *International Conference on Machine Learning.* PMLR, 2803–2813.

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

[18] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. Bayesian modeling of intersectional fairness: The variance of bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining.* SIAM, 424–432.

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

[20] James R Foulds and Shimei Pan. 2020. Are Parity-Based Notions of AI Fairness Desirable? *Bulletin of the IEEE Technical Committee on Data Engineering* 43, 4 (2020), 51–73.

[21] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems.* 3315–3323.

[22] John E Hunter and Frank L Schmidt. 1976. Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin* 83, 6 (1976), 1053.

[23] Rashidul Islam, Kamrun Naher Keya, Shimei Pan, and James R Foulds. 2019. Mitigating demographic biases in social media-based recommender systems. *In the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Social Impact Track)* (2019).

[24] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James R Foulds. 2021. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of The Web Conference 2021.*

[25] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press.

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning.* PMLR, 2564–2572.

[27] Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James R Foulds. 2021. Equitable allocation of healthcare resources with fair survival models. In *Proceedings of the 2021 SIAM International Conference on Data Mining.* SIAM.

[28] Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations.* 1–15.

[29] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems.*

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[31] Gilles Louppe, Michael Kagan, and Kyle Cranmer. 2017. Learning to pivot with adversarial networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems.* 982–991.

[32] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. 2020. There is no trade-off: enforcing fairness can improve accuracy. *arXiv preprint arXiv:2011.03173* (2020).

[33] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency.* 107–118.

[34] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31.

[35] Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James R Foulds. 2020. Equitable Allocation of Healthcare Resources with Fair Cox Models. In *AAAI Fall Symposium on AI in Government and Public Sector.* AAAI FSS.

[36] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

[37] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems (Autodiff Workshop).*

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 32 (2019), 8026–8037.

[40] Valerio Perrone, Michele Donini, Krishnaram Kenthapadi, and Cédric Archambeau. 2020. Bayesian optimization with fairness constraints. *International Conference on Machine Learning (Automated Machine Learning Workshop)* (2020).

[41] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[42] Jasper Snoek, Hugo Larochelle, and Ryan Prescott Adams. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems* (2012).

[43] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics.* PMLR, 2164–2173.

[44] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning.* 6373–6382.

[45] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[46] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.* 20, 75 (2019), 1–42.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics.* 962–970.

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

[49] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 335–340.

[50] Han Zhao and Geoff Gordon. 2019. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems.* 15675–15685.