Accurate Fairness: Improving Individual Fairness without Trading Accuracy

Xuran Li [1 2] **Peng Wu** [1 2] **Jing Su** [1 2]

Abstract

Accuracy and individual fairness are both crucial aspects for trustworthy machine learning. However, they are incompatible with each other, and enhance one aspect may sacrifice the other inevitably. We propose in this paper a new fairness criteria, *accurate fairness*, to assess whether an individual is treated both accurately and fairly regardless of protected attributes. Thus, the side effects of enhancing just one of the two aspects, i.e., true bias and false fairness, can be effectively identified with our criteria. We then present a *Siamese Fairness* approach for accurate fairness training. To the best of our knowledge, this is the first time that the Siamese network is adapted for bias mitigation. Case studies with typical fairness benchmarks demonstrate that our fair Siamese approach can, on average, promote the 17.4% higher individual fairness, the 11.5% higher fair-F1 score, and the 4.7% higher accuracy of a machine learning model than the state-of-the-art bias mitigation techniques. Finally, our approach is applied to mitigate the possible service discrimination with a real Ctrip dataset, by fairly serving on average 97.9% customers with different consumption habits who pay the same prices for the same rooms (20.7% more than original models).

1. Introduction

Machine learning-based intelligent systems have exhibited competitive performances, in terms of their accuracy and efficiency, for real-world decision making tasks, e.g., loan granting (Hardt et al., 2016), criminal justice risk assessment (Berk et al., 2021), and online recommendations (Lambrecht & Tucker, 2019). While the public enjoys the benefit of such technology advances in general, the widespread deployments of these machine learning systems have also spawned social and ethnic concerns on individuals' daily lives, particularly regarding the fairness of the decisions or predictions made by the machine learning systems.

Accuracy and fairness are both crucial aspects for trustworthy machine learning. Informally, accuracy assesses how well a machine learning system fits the ground truth distribution, while fairness assesses the decisions or predictions of a machine learning system from the perspective of social and ethnic equity. However, they are incompatible with each other and enhance one aspect may sacrifice the other inevitably with unacceptable consequences (Dutta et al., 2020; Kim et al., 2020). For instance, more accurate predictions on applicants' incomes benefit banks with less lending risk. But historical loan data may contain social discrimination regarding protected attributes such as genders, races, and ages. Thus, on one hand, accurate predictions of a machine learning system, trained with the historical loan data, would reflect, even exaggerate such discrimination against some individuals. On the other hand, enhancing just its fairness, e.g., individual fairness (Dwork et al., 2012; Galhotra et al., 2017), by blindly enforcing the applicants with different genders, races or ages to have the same access to loans, would result in more faulty predictions that are fair but inconsistent with the ground truths, hence harm the reliability of the machine learning system. Therefore, accurate but biased, and fair but faulty predictions do not yield a mutually beneficial trade-off between accuracy and fairness.

In this paper we propose to investigate the individual fairness of a machine learning system in alignment with its accuracy, in order to deliver truthfully fair solutions for real-world decision making tasks. In particular, we focus on individual fairness, which may induce collectively group fairness (e.g., statistical parity (Calders et al., 2009; Kamiran & Calders, 2009), confusion matrix-based fairness (Caton & Haas, 2020)), but not vice versa generally. we came from the intuition that every individual should be treated accurately and should not be treated differently because of their sensitive protected attributes, e.g. gender, race, age, etc, as acknowledging the differences by sensitive attributes automatically introduces difference, conflict and bias (Gündemir & Galinsky, 2017).

we integrate the notion of individual fairness with the correctness constraint from the notion of accuracy to propose

[1]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences. Correspondence to: Xuran Li <lixr@ios.ac.cn>, Peng Wu <wp@ios.ac.cn>, Jing Su <sujing@ios.ac.cn>.

accurate fairness. The prediction result for an individual is *accurately fair*, if it is both consist to its ground-truth and independent to its protected attributes; otherwise, the prediction result is either faulty or biased. Accurate fairness captures exactly the intuition that fairness criteria shall be truthfully built upon accurate predictions, while fair but faulty predictions fabricate a false fairness effect that actually conceals individual discrimination instead of eliminating it, and accurate but biased predictions make such discrimination persist, if not worsen. We further propose fairness confusion matrix to identify the incompatibility between accuracy and individual fairness, and new fairness metrics, *fair-precision*, *fair-recall* and *fair-F1 score* to evaluate the reliability of a machine learning model. Fair-precision is the proportion of the individually fair predictions in the accurate predictions. Fair-recall is the proportion of the accurate predictions in the individually fair predictions. Fair-F1 score is the harmonic mean of fair-precision and fair-recall.

In addition, we present and implement a Siamese fairness approach to train a machine learning model (e.g., a logistic regression model or a neural network model) with individual inputs and their similar counterparts, under the accurate fairness constraints, to mitigate individual bias without trading its accuracy. Empirical studies with typical fairness benchmarks *Adult (Census Income)* (Adult, 1996), *German Credit* (Credit, 1994), and *ProPublica Recidivism (COMPAS)* (Angwin et al., 2016) demonstrate that the accurate fairness notion contributes well to accuracy and individual fairness simultaneously. Compared with the state-of-the-art bias mitigation techniques, our approach can on average promote the 17.4% higher individual fairness and the 4.7% higher accuracy of a machine learning model, with the 11.5% higher fair-F1 score in terms of accurate fairness.

Finally, we apply the accurate fairness metric to evaluate a service discrimination problem with a real dataset (Heywhale, 2019) from Ctrip, one of the largest online travel service providers in the world. This problem concerns whether customers who pay the same prices for the same rooms are recommended with the same room services, irrespective of their consumption habits. Discrimination in customer services has gained much attention, especially when nowadays services are more often provided through algorithmic recommendations. We train three neural network models with the Ctrip dataset, which do suffer service discrimination against the customers with different consumption habits. Our approach can mitigate the service discrimination to a great extent, by fairly serving on average 97.9% customers with different consumption habits who pay the same prices for the same rooms (20.7% more than the original models).

The main contributions of this paper are as follows.

- We propose a new fairness metric, *accurate fairness*, which is an organic synergy between accuracy and individual fairness, requiring that individuals be treated both accurately and fairly regardless of their protected attributes.
- We propose fairness confusion matrix to identify the incompatibility between accuracy and individual fairness,and accurate fairness metrics, *fair-precision, fair-recall and fair-F1 score*, to evaluate the level of accurate fairness of a machine learning model.
- We present and implement a Siamese fairness approach to train a machine learning model under the metric of accurate fairness, in order to improve its individual fairness without trading its accuracy. To the best of our knowledge, this is the first time that a Siamese approach is adapted for individual bias mitigation.
- The accurate fairness and fair Siamese approach are applied with typical fairness benchmarks and a real Ctrip dataset. The case studies reveal the existence of true bias and false fairness in the learned classifiers, while our approach can truthfully mitigate these defects to a satisfying extent.

The rest of this paper is organized as follows. We briefly discuss the related work in Section 2, followed by the presentation and demonstration of the accurate fairness notion in Section 3. We present the fair Siamese approach in Section 4. Its implementation and evaluation results are reported and analyzed in Section 5. The paper is concluded in Section 6 with some future work.

2. Related Work

2.1. Fairness Criteria

The fairness criteria presented in literature are usually partitioned into two categories: group fairness criteria and individual fairness criteria.

Group fairness criteria concern equal treatments for the groups of the individuals with the same protected attribute values, while ignoring their other attributes. Thus, the group fairness criteria are usually defined statistically in terms of conditional independence. Statistical parity (Calders et al., 2009; Kamiran & Calders, 2009; 2012) requires the same positive prediction rate for each group of the individuals with the same protected attribute values, so that the predictions are independent of the protected attributes. Ground truths can also join in the statistical fairness metrics, yielding confusion matrix-based fairness (e.g., equality odds (Hardt et al., 2016), accuracy equality (Berk et al., 2021)) and calibration-based fairness (e.g., calibration fairness (Pleiss et al., 2017; Chouldechova, 2017)). The former implies that the predictions are independent of the protected attributes under the given ground truths, while the latter implies that

the ground truths are independent of the protected attributes under the given predictions. However, the group fairness focuses on the statistical treatment equal between groups, which neglects the fairness of individual, resulting in individuals are unfavorably discriminated in contrast to their similar counterparts (Makhlouf et al., 2021).

Individual fairness criteria can be defined qualitatively or quantitatively by interpreting the notions of *similar individuals* and *similar treatments*, in order to assess whether similar individuals are treated similarly. Causal discrimination (Galhotra et al., 2017) is such a qualitative definition, where similar individuals are those who differ only on the protected attributes, and only the equal predictions are accounted similar treatments. In a quantitative or algorithmic definition, task-specific distance metrics are involved to characterize the similarities between individuals and between prediction distributions. Individual fairness (Dwork et al., 2012) requires that the similarity distance between individuals lays an upper bound on the similarity distance between the corresponding prediction distributions. Moreover, counterfactual fairness (Kusner et al., 2017) takes into account task-specific causalities between attributes and predictions, and focuses on the equality of the prediction distributions in the actual and counterfactual worlds.

Especially, the individual fairness criteria directly judge the fairness of the predictions themselves (Xie & Wu, 2020), which may lead to false fairness achieved based on faulty predictions. Accurate fairness presented in this paper refactors the individual fairness criteria from a viewpoint of accuracy, so that mitigating individual bias of a machine learning model does not necessarily sacrifice its accuracy, improving the individual fairness criteria in a sensible and plausible way. Please refer to (Galhotra et al., 2017; Dwork et al., 2012; Kusner et al., 2017; Caton & Haas, 2020; Makhlouf et al., 2021; Berk et al., 2021) for a comprehensive survey about machine learning fairness notions.

2.2. Bias Mitigation

Bias of a machine learning model can be mitigated through pre-processing the training data, in-processing the model itself or post-processing the predictions, as summarized in (Caton & Haas, 2020; Bellamy et al., 2019).

Pre-processing techniques mitigate the bias in the training data. Among others, a reweighing algorithm (Calders et al., 2009; Kamiran & Calders, 2012) learns different weights for different groups divided by the protected attributes and the ground truths, to construct a balanced dataset. A fair representation learning approach (Zemel et al., 2013) intends to learn an intermediate representation of the training data to obfuscate any information about their protected attributes. Another process, iFair (Lahoti et al., 2019) learns a generalized data representation preserving the fairness-aware similarity between individuals and minimizing the data loss to reconcile individual fairness with the utility of model. Disparate impact remover (Feldman et al., 2015) changes the values of none-protected to preserve the relative per-attribute rank in the original dataset and construct a no disparate impact dataset. Recently, an approach of removing biased data (Verma et al., 2021) was proposed to use an influence function to identify and remove directly the biased data for model retraining.

In-processing techniques train a machine learning model with fairness as an additional optimization goal. For instance, an adversarial debiasing approach (Zhang et al., 2018) maximizes the model's prediction accuracy and in the meanwhile minimizes the likelihood of an adversary to predict the individuals' protected attributes from the predictions. Meta fair classifier (Celis et al., 2019) produces an approximately fair model through a new-meta algorithm for classification, which takes as input any a general class of fairness constraints phrased as "linear-fractional constraints". Gerry fair classifier (Kearns et al., 2018; 2019) based on the equilibrium of the game proposes a two-player zero-sum game between a Learner and an Auditor , that the Learner play the no-regret Follow the Perturbed Leader algorithm and the Auditor play best response, which converges to an approximate Nash equilibrium and the best fair distribution. Sensitive set invariance (Yurochkin & Sun, 2021) develops a stochastic approximation algorithm to minimize the transport-based regularizer enforcing the treatments invariance on certain sensitive sets, to achieve distributional individual fairness. Sensitive subspace robustness (Yurochkin et al., 2020) develops a distributionally robust optimization approach to enforce the robustness to certain sensitive perturbations to the inputs during training

Post-processing techniques mitigate the bias in the predictions through relabelling. The equalized odds (Hardt et al., 2016) or calibrated equalized odds (Pleiss et al., 2017) algorithms compute the optimal conditional probabilities of relabelling with the accuracy loss minimized under the equalized odds or calibrated equalized odds constraints. Based on the prediction confidence and ensemble disagreement, reject option classifier (Kamiran et al., 2012) exploits the low confidence region of a probabilistic classifiers to invoke the reject option and label instances belonging to deprived and favored groups to reduce discrimination.

These bias mitigation processing could mitigate the bias on the basis of a given fairness criterion, but at a cost of accuracy loss. The inconsistency fairness criterion and the incompatibility between accuracy and fairness bring challenges for evaluate the bias mitigation efficiency of the processing. Accurate independence is the first unified fairness criteria which also compatible with the accuracy, so that our criterion could quantity the trade-off between accuracy and

fairness and analysis the changes of individual fairness and group fairness changings after mitigation.

3. Accurate fairness

We present in this section the notion of accurate fairness and discuss its connection with group fairness.

Assume a finite and labelled dataset D with the domains of the protected attributes, the remaining (i.e., non-protected) attributes, and the ground truth labels denoted A, X, Y, respectively. Each input $v \in D$ is represented as a tuple $v = (x, a)$ with $x \in X, a \in A$, and associated with a ground truth label $y \in Y$. $v = (x, a)$ forms group of individuals who have the same none-protected attributes but different protected attributes, referred to as the *accurately fair group of* v .

$$I(x, a) = \{(x, a') \mid \forall a' \in A\}$$

Let $f : X \times A \to \hat{Y}$ denote a classifier learned based on the dataset D, and $\hat{y} = f(x, a)$ the prediction result of classifier f for input $v = (x, a)$.

Definition 3.1 (Accurate Fairness)**.** A classifier $f : X \times A \to \hat{Y}$ is *accurately fair* to input $v = (x, a) \in D$, if $\forall (x, a'), (x, a'') \in I(x, a)$

$$\begin{aligned} D(f(x, a'), f(x, a'')) &\leq Ld((x, a'), (x, a'') \\ D(y, f(x, a'))) \quad &\leq Ld((x, a), (x, a') \end{aligned}$$

where y is the ground truth for v, $f(x, a'), f(x, a'')$ is the predication for the individual in the accurately fair group of v

The above definition requires that the individuals in the accurately fair group of $v = (x, a)$ should get the equal treatment (i.e., $\forall (x, a'), (x, a'') \in I(x, a), D(f(x, a'), f(x, a'')) \leq Ld((x, a'), (x, a''))$ and all of them should be consist to the ground-truth of v (i.e., $\forall (x, a') \in I(x, a), D(y, f(x, a'))) \leq Ld((x, a), (x, a'))$. Thus, it refines the notion of individual fairness with the notion of accuracy, to enforce *truthful* individual fairness based on accurate predictions. Herein, accurately fair group of v get only one label y of v, as other similar counterparts may not exist in D.

The accurate fairness notion further induces a fairness confusion matrix, as shown in Table 1, summarizing the orthogonal synergy between individual fairness and accuracy to identify the incompatibility between accuracy and individual fairness.

Definition 3.2 (Fairness Confusion Matrix)**.** For classifier $f : X \times A \to \hat{Y}$ and input $(x, a) \in D$, the prediction result $f(x, a)$ is *true fair* if it is both correct and individually fair; $f(x, a)$ is *true biased* if it is correct but individually biased; $f(x, a)$ is *false fair* if it individually fair but not correct; $f(x, a)$ is *false biased* if it is neither correct, nor individually fair. Thus, the following new metrics: *True Fair Rate (TFR)*,*True Biased Rate (TBR)*, *False Fair Rate (FFR)*, *False Biased Rate (FBR)*, *Fair-Precision (F-P)*, *Fair-Recall (F-R)*, and *Fair-F1 Score (F-F1)* can be defined with respect to the dataset D.

Table 1. Fairness Confusion Matrix

Accuracy \ Fairness	Fair	Biased
True	True Fair	True Biased
False	False Fair	False Biased

$$TFR = \frac{1}{n} \sum_{(x,a)\in D} \mathbb{I}\Big(\begin{array}{l} y = f(x, a) \text{ and} \\ f(x, a) = f(x, a') \text{ for any } a' \neq a \end{array} \Big)$$

$$TBR = \frac{1}{n} \sum_{(x,a)\in D} \mathbb{I}\Big(\begin{array}{l} y = f(x, a) \text{ and} \\ f(x, a) \neq f(x, a') \text{ for some } a' \neq a \end{array} \Big)$$

$$FFR = \frac{1}{n} \sum_{(x,a)\in D} \mathbb{I}\Big(\begin{array}{l} y \neq f(x, a) \text{ and} \\ f(x, a) = f(x, a') \text{ for any } a' \neq a \end{array} \Big)$$

$$FBR = \frac{1}{n} \sum_{(x,a)\in D} \mathbb{I}\Big(\begin{array}{l} y \neq f(x, a) \text{ and} \\ f(x, a) \neq f(x, a') \text{ for some } a' \neq a \end{array} \Big)$$

$$F\text{-}P = \frac{TFR}{TFR + TBR}$$

$$F\text{-}R = \frac{TFR}{TFR + FFR}$$

$$F\text{-}F1 = \frac{2 \times F\text{-}P \times F\text{-}R}{F\text{-}P + F\text{-}R}$$

where $n = |D|$ is the size of D, and $\mathbb{I}(\cdot)$ is the indicator function such that $\mathbb{I}(\cdot) = 1$ if $\cdot$ holds and 0 otherwise.

Through the fairness confusion matrix, it can be seen that accurate fairness decisively excludes the false fair predictions from individual fairness, and identifies the true biased predictions from accuracy to be further mitigated. The fair-precision metric calculates the individually fair proportion in the accurate predictions, the fair-recall metric calculates the accurate proportion in the individually fair predictions, which measure the compatibility between accuracy and independent fairness from the perspective of accuracy and independence fairness. The fair-F1 score combines fair-precision and fair-recall to measure the level of accurate fairness of a machine learning model comprehensively.

3.1. Connection to Group Fairness

The relations of accurate fairness with accuracy and individual fairness are obvious by definition. We discuss in this section the relationship between accurate fairness and group fairness, particularly statistical parity and confusion

matrix-based fairness.

Accurate fairness endorses groups fairness notions (statistical parity and confusion matrix-based fairness) collectively over the accurately fair groups. Let $I(D') \subseteq X \times A$ denote the union of the all accurately fair groups in D', referred to as the *accurately fair set of* D', i.e.,

$$I(D') = \bigcup_{(x,a)\in D'} I(x,a)$$

By definition, $D' \subseteq I(D')$. Under the group fairness notions, assume each individual $(x,a) \in I(D')$ is associated with the same ground truth label as some $(x,a) \in D'$ is.

Theorem 3.3. *If a classifier $f : X \times A \to \hat{Y}$ is accurately fair to each input in $D' \subseteq D$, then the classifier f satisfies statistical parity and confusion matrix-based fairness over the inter-group similar set $I(D')$.*

Proof. The accurate fairness constraint (3.1) implies that

$$P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = a') \qquad (1)$$

over $I(D')$ for any $a \neq a'$, and hence

$$P(\hat{Y} \neq Y|A = a) = P(\hat{Y} \neq Y|A = a') \qquad (2)$$

over $I(D')$. Thus, the accuracy (or inaccuracy) for the inter-group similar individuals are independent on the protected attributes. As a result, the classifier f satisfies the confusion matrix-based fairness over $I(D')$. Then, with (1) and (2), the classifier f also satisfies statistical parity over $I(D')$, because $P(\hat{Y}|A = a) = P(\hat{Y} = Y|A = a) + P(\hat{Y} \neq Y|A = a)$. □

4. Accurate Fairness In-processing

We present in this section a fair Siamese approach for accurate fairness in-processing, in order to mitigate the true biased and false fair cases demonstrated in the above example. It takes in the pairs of the training data and their similar counterparts to train a machine learning model in a Siamese network (Chopra et al., 2005) to minimize the accuracy loss and the individual fairness loss at the same time. As a by-product of this approach, the individual fairness of the machine learning model can also be improved without trading its accuracy.

A Siamese network provides a mechanism to train identical models with shared parameters. The models receive their own inputs separately in parallel, and jointly produce outputs subject to a shared loss function. It is often applied in metric learning (Kaya & Bilge, 2019) to characterize the similarity between the inputs. Herein, we use a Siamese network to correlate the training processes for the pre-determined similar inputs, so as to guide the identical models to make true fair predictions for the similar inputs, subject to the accurate fairness metric. The architecture of the fair Siamese approach is shown in Figure 1, while its workflow is shown in Algorithm 1.

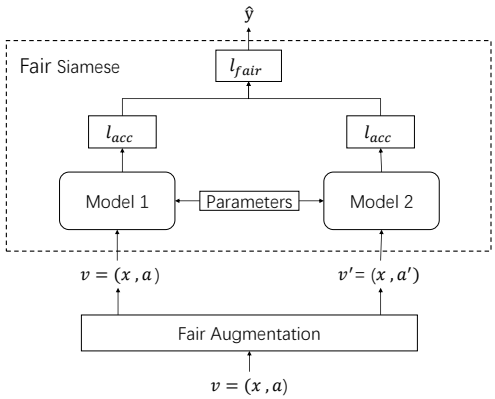


Figure 1. A fair Siamese algorithm

Algorithm 1 (FS)

Input: dataset D, classifier f_θ, learning rate η
Output: classifier f_θ

1: $D^+ \leftarrow \emptyset$;
2: **for** each $v = (x,a) \in D$ **do**
3: $\quad v' \leftarrow (x,a')$ for some $a' \in A$ and $a' \neq a$;
4: $\quad D^+ \leftarrow D^+ \cup \{(v,v')\}$;
5: **end for**
6: Initialize parameters θ;
7: **repeat**
8: $\quad$ **for** each $((x,a),(x,a')) \in D^+$ **do**
9: $\quad\quad \hat{y} \leftarrow f_\theta(x,a)$;
10: $\quad\quad \hat{y}' \leftarrow f_\theta(x,a')$;
11: $\quad\quad$ **for** each parameter $w \in \theta$ **do**
12: $\quad\quad\quad w \leftarrow w - \eta \frac{\partial L_{AF}(y,\hat{y},\hat{y}')}{\partial w}$
13: $\quad\quad$ **end for**
14: $\quad$ **end for**
15: **until** θ converges or the maximal number of iterations is reached
16: **return** f_θ

At Lines 2-4, for each input $v = (x,a) \in D$, its similar counterpart $v' = (x,a')$ is generated with a' randomly chosen in A such that $a' \neq a$. Thus, the labeled dataset D is augmented with the similar counterparts for the subsequent training in a Siamese network. Both models in Figure 1 are the identical copies of classifier f_θ with the shared parameters θ: Model 1 is trained with inputs $(x,a) \in D$, while Model 2 is trained with their similar counterparts (x,a'). Herein, for the sake of evaluation, the number of the training inputs for each model is the same as the size of the original dataset D. In practice, Algorithm 1 can be easily extended to sample multiple similar counterparts for each input in D, particularly for multi-valued protected attributes or a combination of multiple protected attributes.

At Lines 11-13, the shared parameters θ are obtained by applying an error BackPropagation (BP) algorithm (Werbos, 1974; Rummelhart et al., 1986a;b) to the following optimization problem

$$\arg\min_{\theta} \mathbb{E}[L_{AF}(y, f_\theta(x,a), f_\theta(x,a'))]$$

where (x,a) and (x,a') are similar inputs with $a' \neq a$, and y is the ground truth for input (x,a). The accurate fairness loss function $L_{AF}(y,\hat{y},\hat{y}')$ is schematically defined below.

$$L_{AF}(y,\hat{y},\hat{y}') = l_{acc}(y,\hat{y}) + l_{acc}(y,\hat{y}') + l_{fair}(\hat{y},\hat{y}')$$

where $\hat{y} = f(x,a)$ and $\hat{y}' = f(x,a')$ are the prediction results for inputs (x,a) and (x,a'), respectively; $l_{acc}(y,z)$ represents the accuracy loss between the (expected) ground truth y and the prediction result z; and $l_{fair}(z_1,z_2)$ represents the individual fairness loss between the prediction results z_1 and z_2 for similar inputs. In this way, the accuracy requirement is enforced through $l_{acc}(y,\hat{y})$ for input $(x,a) \in D$, and in the meanwhile, the individual fairness requirement is enforced through $l_{fair}(\hat{y},\hat{y}')$ for input $(x,a) \in D$ and its similar counterpart $(x,a') \in X \times A$, and further strengthened with $l_{acc}(y,\hat{y}')$ by explicitly enforcing the conclusion of the accurate fairness constraint (3.1), i.e., the prediction result $\hat{y}'$ for input (x,a') shall also equal the ground truth y for input (x,a).

5. Implementation and Evaluation

We implement the fair Siamese approach (Algorithm 1) in Python 3.8 with TensorFlow 2.6.0. The implementation is evaluated on a Ubuntu 18.04.3 system with Intel Xeon Gold 6154 @3.00GHz CPUs, GeForce RTX 2080 TI GPUs and 512G memory. We use the Mean Squared Error (MSE) loss function as l_{acc} and l_{fair}, which shows the best performance in comparison with other loss functions, e.g., categorical crossentropy or hinge, as reported in Appendix A.1. We use the Adam optimizer (Kingma & Ba, 2015) to adaptively tune the learning rate, which is initialized as 0.001 by default.

The following experiments are designed to investigate the effectiveness of the accurate fairness notion in revealing a decision making model's genuine capability of fairly treating similar individuals, and to investigate the bias mitigation effectiveness of the fair Siamese approach, in comparison with the state-of-the-art bias mitigation techniques, with regard to binary or multi-valued protected attributes, or the combinations thereof.

5.1. Datasets and Models

We use the three benchmark datasets, Adult, German Credit, and COMPAS, which are often used in fairness literature, and a real dataset from Ctrip for the evaluation. The instances with unknown or empty values have been removed from the datasets before training. Table 2 reports the size and the protected attributes of each dataset, and the models trained with these datasets.

Table 2. Datasets and Models

Dataset	Size	Models	Protected Attributes
Adult (Census Income) (Adult, 1996)	45222	FCNN(7) LR	gender (binary) age (multi-valued) race (multi-valued)
German Credit (Credit, 1994)	1000	FCNN(4) LR	gender (binary) age (multi-valued)
ProPublica Recidivism (COMPAS) (Angwin et al., 2016)	6172	FCNN(4) LR	gender (binary) age (multi-valued) race (multi-valued)
Ctrip (Heywhale, 2019)	68191	FCNN(3) FCNN(5) FCNN(7)	customer consumption habits

For each benchmark dataset, a logistic regression (LR) classifier and a fully connected neural network (FCNN) classifier are trained for each of its protected attributes, as well as for their combination. Three FCNN classifiers are trained for the Ctrip dataset. In Table 2, FCNN(l) means an FCNN classifier with l layers. These classifiers are referred to as the baseline (BL) models in the evaluation. Then, for each BL model, let RW, EO, CEO, LFR, AD, RB, FS represent the classifiers resulted by applying the approaches of reweighing, equalized odds, calibrated equalized odds, fair representation learning, and adversarial debiasing implemented in an open-source toolkit AI fairness 360 (Bellamy et al., 2019), the approach of removing biased data (Verma et al., 2021), and our fair Siamese approach, respectively.

5.2. Mitigating Individual Bias

Table 3 reports the average statistics of the accuracy (ACC), individual fairness (IF), TFR, TBR, FFR and FBR metrics for each bias mitigation approach compared over the three benchmark datasets. It can be seen that, among others, our fair Siamese approach can always improve the individual fairness of the BL models to the highest extent without sacrificing their accuracy. Indeed, compared with the other bias mitigation approaches, our fair Siamese approach can on average promote the 17.4% higher individual fairness and the 4.7% higher accuracy of a classifier, with the 11.5% higher fair-F1 score in terms of accurate fairness.

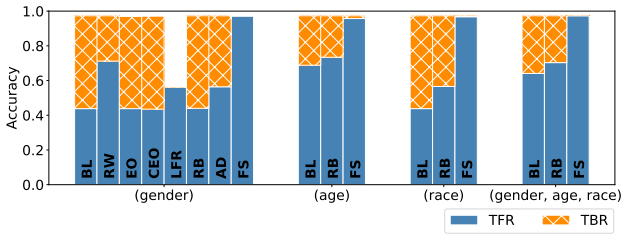
Due to the page limit, we here discuss the accuracy, IF and TFR metrics of the classifiers for the COMPAS dataset. Similar observations can be made for the other benchmark datasets. Please refer to Appendix A.2 for the detailed experimental results of bias mitigation.

As shown in Figure 2, our fair Siamese approach makes the most accurate predictions to be individually fair, while

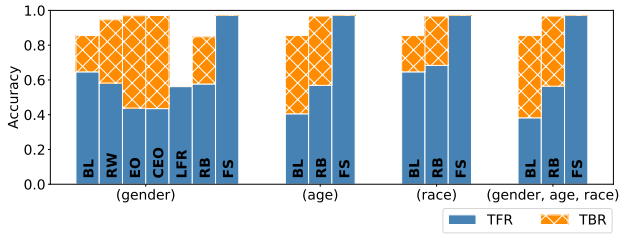
Table 3. Average Statistics over the Three Benchmark Datasets

Model	ACC	IF	TFR	TBR	FFR	FBR	F-P	F-R	F-F1
BL	0.835	0.816	0.675	0.160	0.140	0.024	0.822	0.838	0.809
RW	0.831	0.866	0.715	0.116	0.151	0.018	0.876	0.840	0.840
EO	0.846	0.751	0.628	0.218	0.123	0.032	0.767	0.854	0.778
CEO	0.843	0.737	0.615	0.227	0.121	0.036	0.756	0.854	0.772
LFR	0.671	**1.000**	0.671	0.000	0.329	0.000	**1.000**	0.671	0.801
RB	0.845	0.827	0.696	0.150	0.131	0.024	0.839	0.854	0.828
AD	0.837	0.735	0.647	0.190	0.088	0.076	0.788	**0.891**	0.820
FS	**0.859**	0.993	**0.856**	0.004	0.138	0.003	0.995	0.861	**0.921**

maintaining the original (BL) models' accuracy, hence significantly reduce the most true biased predictions.

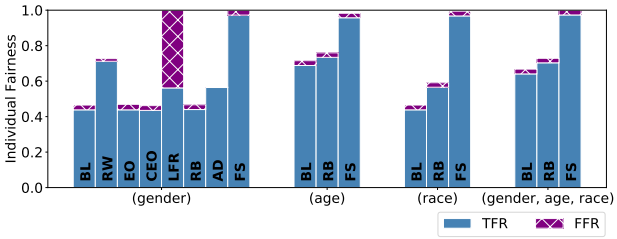


(a) FCNN classifiers for COMPAS

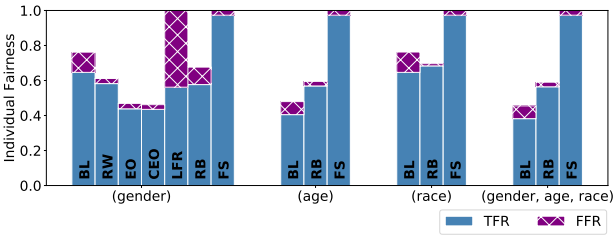


(b) LR classifiers for COMPAS

Figure 2. TFR and Accuracy



(a) FCNN classifiers for COMPAS



(b) LR classifiers for COMPAS

Figure 3. TFR and IF

As shown in Figure 3, our fair Siamese approach also achieves almost the highest level of individual fairness, while making the most individually fair predictions to be accurate. The LFR models, i.e., the classifiers trained through the approach of learning fair representations for the sake of individual fairness, achieve the highest 100 percent of individual fairness (IF=1.0), but about 43.8 percent of all the (individually fair) predictions are not accurate (FFR=0.438). Through examining their predication results, we find that this is because the LFR models make the same predictions for all the inputs. Hence, one-sided enforcement of individual fairness may result in false fairness with a significantly negative impact on prediction accuracy, which is exactly the notion of accurate fairness intends to expose.

5.3. Service Discrimination with the Ctrip Dataset

We then apply the accurate fairness notion and the fair Siamese approach to investigate a service discrimination problem, where customers with different consumption habits may be recommended with disparate services, even though they pay the same prices for the same rooms. The Ctrip dataset includes six consumption habit attributes C1, ..., C6 of customers (representing the average time of order confirmation, the average days of advance booking, the average star level, class level, recommended level of hotels booked, and the average days of hotel stay, respectively) and six attributes H1, ..., H6 of hotels (representing order date, hotel ID, room type, room ID, star level and room price, respectively). For the service discrimination problem, the six customer attributes are designated as the protected attributes. The ground truth labels represent the room service types.

As reported in table 4, on average, only 55.8% (TFR) customers are treated accurately and fairly by the original (BL) models. Through fair Siame in-processing, not only their average TFR is improved to 66.0% under the upper bound of their average accuracy, which is also improved to 67.0%. Our fair Siamese approach can make most (on average 97.9%) of the customers to be fairly served irrespective of their consumption habits, as long as they pay the same prices for the same rooms. Thus, to further improve the in-

dividual fairness truthfully, it is left to improve the accuracy of the classifiers themselves, instead of sacrificing it.

Table 4. Ctrip

Model	Acc		IF		TFR		F-F1	
	BL	FS	BL	FS	BL	FS	BL	FS
FCNN(3)	0.663	0.665	0.746	0.989	0.547	0.660	0.777	0.798
FCNN(5)	0.670	0.679	0.807	0.971	0.573	0.665	0.776	0.807
FCNN(7)	0.667	0.666	0.764	0.976	0.554	0.656	0.775	0.798
average	0.667	**0.670**	0.772	**0.979**	0.558	**0.660**	0.776	**0.801**

6. Conclusion

We present in this paper the accurate fairness notion that sheds a new light on studying individual fairness from the perspective of accuracy. It is built upon the conditional independence constraint from the notion of individual fairness, in conjunction with the correctness constraint from the notion of accuracy. The accurate fairness notion sets up the fairness confusion matrix that can clarify the side effects of trading accuracy for individual fairness and vice versa: accurate but individually biased predictions reflect the latent discrimination in training data or decision making models themselves; while individually fair but faulty predictions reveal the faked fairness effect that a decision making model achieves with sacrifice of its accuracy. We further propose the new metrics of fair-precision, fair-recall and fair-F1 score to systematically evaluate the reliability of a decision making model from the perspective of accurate fairness. Then we present and evaluate the fair Siamese in-processing approach to train the decision making model for the sake of accurate fairness, which also significantly improves its individual fairness without trading its accuracy.

Most work on individual fairness rely on a pre-specified setting of protected attributes and disadvantaged groups. As part of future work, the fairness confusion matrix can be adapted to identify which sensitive attributes poss more impacts on prediction outcomes. The accurate fairness notion can be utilized to help further diagnose which groups under these attributes are treated unfavorably.

References

Adult. Census Income dataset. UCI machine learning repository. https://archive-beta.ics.uci.edu/ml/datasets/adult, 1996.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. in ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, pp. 4:1–4:15, 2019.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 3–44, 2021.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009*, pp. 13–18, 2009.

Caton, S. and Haas, C. Fairness in machine learning: A survey. *CoRR*, 2020.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 319–328. ACM, 2019.

Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pp. 539–546, 2005.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, pp. 153–163, 2017.

Credit. German Credit Data. UCI machine learning repository. https://archive-beta.ics.uci.edu/ml/datasets/statlog+german+credit+data, 1994.

Dutta, S., Wei, D., Yueksel, H., Chen, P., Liu, S., and Varshney, K. R. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pp. 2803–2813, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pp. 214–226, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pp. 259–268. ACM, 2015.

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

Gündemir, S. and Galinsky, A. D. Multicolored blindfolds: How organizational multiculturalism can conceal racial discrimination and delegitimize racial discrimination claims. *Social Psychological and Personality Science*, pp. 194855061772683, 2017.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016.

Heywhale. https://www.heywhale.com/mw/project/5ca2d6098408c1002b48bf3c/dataset, 2019. in Chinese.

Kamiran, F. and Calders, T. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pp. 1–6, 2009.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, pp. 1–33, 2012.

Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929, 2012. doi: 10.1109/ICDM.2012.45.

Kaya, M. and Bilge, H. S. Deep metric learning: A survey. *Symmetry*, pp. 1066, 2019.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2564–2572. PMLR, 2018.

Kearns, M. J., Neel, S., Roth, A., and Wu, Z. S. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 100–109. ACM, 2019.

Kim, J. S., Chen, J., and Talwalkar, A. FACT: A diagnostic for group fairness trade-offs. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pp. 5264–5274, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4066–4076, 2017.

Lahoti, P., Gummadi, K. P., and Weikum, G. ifair: Learning individually fair data representations for algorithmic decision making. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pp. 1334–1345. IEEE, 2019.

Lambrecht, A. and Tucker, C. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 2019.

Makhlouf, K., Zhioua, S., and Palamidessi, C. Machine learning fairness notions: Bridging the gap with real-world applications. *Inf. Process. Manag.*, pp. 102642, 2021.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5680–5689, 2017.

Rummelhart, D., Hinton, G., and Williams, R. Learning internal representations by error propagation. *Nature*, pp. 318–362, 1986a.

Rummelhart, D., Hinton, G. E., and Williams, R. J. Learning representations by back propagating errors. *Nature*, pp. 533–536, 1986b.

Verma, S., Ernst, M. D., and Just, R. Removing biased data to improve fairness and accuracy. *CoRR*, 2021.

Werbos, P. J. Beyond regression: New tools for prediction and analysis in the behavioral science. thesis (ph.d.). appl. math. harvard university. *Ph.D. thesis, Harvard University*, 1974.

Xie, W. and Wu, P. Fairness testing of machine learning models using deep reinforcement learning. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 121–128, 2020.

Yurochkin, M. and Sun, Y. Sensei: Sensitive set invariance for enforcing individual fairness. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Yurochkin, M., Bower, A., and Sun, Y. Training individually fair ML models with sensitive subspace robustness. In *8th*

International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 325–333, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pp. 335–340, 2018.

A. Appendix

A.1. Accurate Fairness Loss Function L_{AF}

Table 5 reports the performance, in terms of TFR, fair-precision (F-P), fair-recall (F-R) and fair-F1 score (F-F1), of the categorical crossentropy (CC), mean squared error (MSE), and hinge (H) loss functions on the three benchmark datasets under various settings of protected attributes. It can be seen that the MSE loss function exhibits the highest average fair-F1 score (0.921). Therefore, we use the MSE loss function in the evaluation experiments.

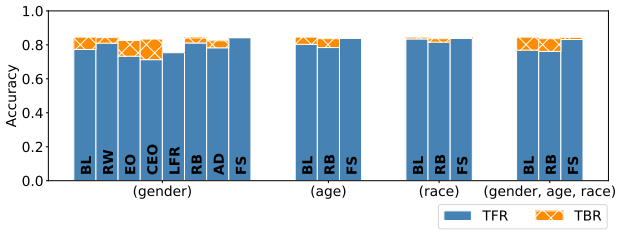
Table 5. Accurate Fairness Loss Function L_{AF} (with CC=Categorical Crossentropy, MSE=Mean Squared Error and H=Hinge)

Dataset	Attr	Model	TFR			F-P			F-R			F-F1		
			CC	MSE	H	CC	MSE	H	CC	MSE	H	CC	MSE	H
Adult	gender	FCNN	0.840	0.842	0.842	0.992	0.996	0.998	0.850	0.848	0.843	0.916	0.916	0.914
	age		0.836	0.838	0.754	0.995	0.994	1.000	0.843	0.846	0.754	0.913	0.914	0.860
	race		0.835	0.839	0.837	0.990	0.998	0.999	0.848	0.844	0.839	0.913	0.914	0.912
	(age, race, gender)		0.837	0.832	0.839	0.992	0.985	0.999	0.847	0.851	0.840	0.913	0.913	0.913
COMPAS	gender		0.973	0.972	0.972	1.000	0.999	0.999	0.973	0.973	0.973	0.986	0.986	0.986
	age		0.973	0.957	0.973	1.000	0.984	1.000	0.973	0.974	0.973	0.986	0.979	0.986
	race		0.969	0.968	0.971	0.996	0.994	0.998	0.973	0.973	0.973	0.984	0.984	0.985
	(age, race, gender)		0.955	0.973	0.972	0.981	1.000	0.999	0.973	0.973	0.973	0.977	0.986	0.986
Credit	gender		0.733	0.743	0.698	1.000	0.993	1.000	0.740	0.754	0.698	0.851	0.857	0.822
	age		0.698	0.753	0.698	1.000	1.000	1.000	0.698	0.760	0.698	0.822	0.864	0.822
	(age, gender)		0.693	0.738	0.698	0.993	0.974	1.000	0.697	0.760	0.698	0.819	0.854	0.822
Adult	gender	LR	0.754	0.829	0.754	1.000	0.997	1.000	0.754	0.834	0.754	0.860	0.908	0.860
	age		0.754	0.832	0.754	1.000	1.000	1.000	0.754	0.833	0.754	0.860	0.909	0.860
	race		0.754	0.829	0.754	1.000	0.997	1.000	0.754	0.833	0.754	0.860	0.908	0.860
	(age, race, gender)		0.754	0.827	0.754	1.000	0.996	1.000	0.754	0.834	0.754	0.860	0.908	0.860
COMPAS	gender		0.562	0.973	0.962	1.000	1.000	0.989	0.562	0.973	0.973	0.719	0.986	0.981
	age		0.562	0.973	0.970	1.000	1.000	0.997	0.562	0.973	0.973	0.719	0.986	0.985
	race		0.562	0.973	0.957	1.000	1.000	0.984	0.562	0.973	0.974	0.719	0.986	0.979
	(age, race, gender)		0.562	0.973	0.972	1.000	1.000	0.999	0.563	0.973	0.975	0.720	0.986	0.987
Credit	gender		0.302	0.757	0.698	1.000	1.000	1.000	0.302	0.757	0.698	0.464	0.862	0.822
	age		0.302	0.723	0.698	1.000	1.000	1.000	0.302	0.723	0.698	0.464	0.839	0.822
	(age, gender)		0.302	0.703	0.698	1.000	1.000	1.000	0.302	0.703	0.698	0.464	0.826	0.822
average			0.705	**0.857**	0.828	0.997	0.996	**0.998**	0.708	**0.862**	0.830	0.809	**0.921**	0.902

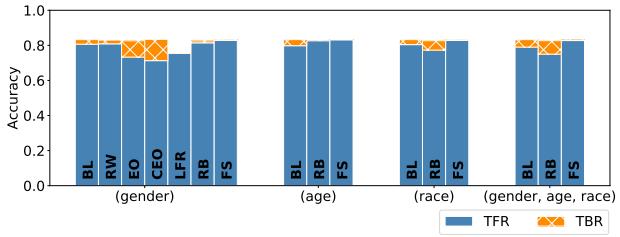
A.2. Evaluation on Each Benchmark Datasets

We present in this section the full evaluation results on the *Adult (Census Income)* , *German Credit*, and *ProPublica Recidivism (COMPAS)* datasets.

A.2.1. ADULT

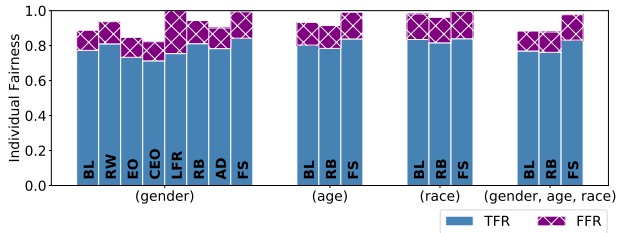


(a) FCNN classifiers for Adult

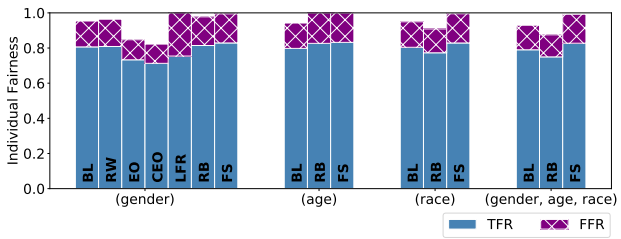


(b) LR classifiers for Adult

Figure 4. TFR and Accuracy



(a) FCNN classifiers for Adult



(b) LR classifiers for Adult

Figure 5. TFR and IF

Table 6. Statistics of FCNN and LR Classifiers for Adult

Attr	Model	FCNN									LR								
		Acc	IF	TFR	TBR	FFR	FBR	F-P	F-R	F-F1	Acc	IF	TFR	TBR	FFR	FBR	F-P	F-R	F-F1
gender	BL	0.845	0.888	0.773	0.073	0.115	0.040	0.914	0.871	0.892	0.833	0.954	0.807	0.026	0.147	0.020	0.968	0.846	0.903
	RW	0.843	0.937	0.810	0.033	0.127	0.031	0.961	0.865	0.910	0.830	0.964	0.809	0.021	0.155	0.015	0.974	0.839	0.902
	EO	0.826	0.847	0.733	0.093	0.114	0.060	0.888	0.865	0.876	0.826	0.847	0.733	0.093	0.114	0.060	0.888	0.865	0.876
	CEO	0.834	0.822	0.713	0.121	0.109	0.056	0.855	0.867	0.861	0.834	0.822	0.713	0.121	0.109	0.056	0.855	0.867	0.861
	LFR	0.754	1.000	0.754	0.000	0.246	0.000	1.000	0.754	0.860	0.754	1.000	0.754	0.000	0.246	0.000	1.000	0.754	0.860
	RB	0.841	0.943	0.811	0.030	0.132	0.027	0.965	0.860	0.909	0.825	0.978	0.814	0.011	0.163	0.012	0.987	0.833	0.904
	AD	0.824	0.901	0.783	0.041	0.118	0.058	0.950	0.869	0.908					-				
	FS	0.846	0.993	**0.842**	0.003	0.151	0.004	0.996	0.848	**0.916**	0.831	0.994	**0.829**	0.002	0.165	0.004	0.997	0.834	**0.908**
age	BL	0.845	0.933	0.803	0.042	0.130	0.025	0.950	0.861	0.903	0.833	0.942	0.798	0.035	0.144	0.023	0.958	0.847	0.899
	RB	0.836	0.915	0.784	0.052	0.131	0.033	0.938	0.857	0.896	0.827	1.000	0.827	0.000	0.173	0.000	1.000	0.827	0.905
	FS	0.843	0.991	**0.838**	0.005	0.153	0.004	0.994	0.846	**0.914**	0.833	1.000	**0.832**	0.000	0.167	0.000	1.000	0.833	**0.909**
race	BL	0.845	0.981	0.835	0.010	0.146	0.009	0.988	0.851	0.914	0.833	0.949	0.804	0.029	0.145	0.022	0.966	0.847	0.903
	RB	0.836	0.962	0.816	0.021	0.146	0.018	0.976	0.848	0.908	0.827	0.909	0.773	0.054	0.136	0.038	0.935	0.851	0.891
	FS	0.841	0.995	**0.839**	0.002	0.156	0.003	0.998	0.844	**0.914**	0.832	0.995	**0.829**	0.003	0.166	0.002	0.997	0.833	**0.908**
(age race gender)	BL	0.845	0.883	0.769	0.076	0.114	0.041	0.910	0.871	0.890	0.833	0.930	0.790	0.043	0.139	0.028	0.949	0.850	0.897
	RB	0.836	0.878	0.762	0.075	0.116	0.047	0.910	0.867	0.888	0.827	0.875	0.750	0.077	0.125	0.048	0.907	0.857	0.881
	FS	0.844	0.977	**0.832**	0.013	0.145	0.011	0.985	0.851	**0.913**	0.831	0.992	**0.827**	0.004	0.165	0.004	0.996	0.834	**0.908**

It can be seen in Table 6 that our fair Siamese approach achieves the highest fair-F1 scores for the FCNN and LR classifiers on the Adult dataset under various setting of protected attributes. It improves the individual fairness of the BL models to the highest extent without sacrificing their accuracy. The '-' means the corresponding approach (adversarial debiasing, AD) is not applicable for the LR classifiers. The observations on true bias and false fairness, discussed in Section 5.2, can be made similarly for the Adult dataset, as shown in Figure 4 and Figure 5.

A.2.2. COMPAS

It can be seen in Table 7 that our fair Siamese approach achieves the highest fair-F1 scores for the FCNN and LR classifiers on the COMPAS dataset under various setting of protected attributes. It improves the individual fairness of the BL models to

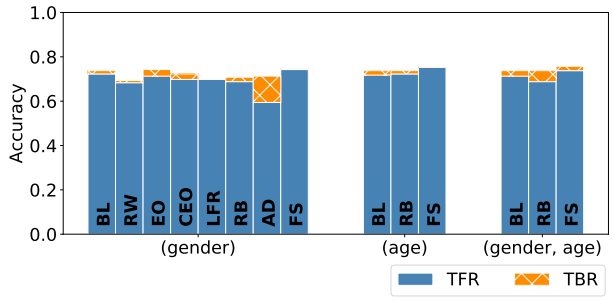
Table 7. Statistics of FCNN and LR Classifiers for COMPAS

Attr	Model	FCNN									LR								
		Acc	IF	TFR	TBR	FFR	FBR	F-P	F-R	F-F1	Acc	IF	TFR	TBR	FFR	FBR	F-P	F-R	F-F1
gender	BL	0.973	0.465	0.438	0.535	0.027	0.000	0.450	0.942	0.610	0.855	0.762	0.646	0.209	0.116	0.029	0.756	0.848	0.799
	RW	0.973	0.727	0.712	0.261	0.015	0.012	0.732	0.979	0.838	0.946	0.612	0.582	0.364	0.029	0.025	0.615	0.952	0.748
	EO	0.969	0.469	0.438	0.531	0.031	0.000	0.452	0.934	0.610	0.969	0.469	0.438	0.531	0.031	0.000	0.452	0.934	0.610
	CEO	0.971	0.462	0.435	0.536	0.027	0.002	0.448	0.942	0.607	0.971	0.462	0.435	0.536	0.027	0.002	0.448	0.942	0.607
	LFR	0.562	1.000	0.562	0.000	0.438	0.000	1.000	0.562	0.719	0.562	1.000	0.562	0.000	0.438	0.000	1.000	0.562	0.719
	RB	0.973	0.467	0.440	0.533	0.027	0.000	0.452	0.942	0.611	0.848	0.676	0.577	0.271	0.100	0.052	0.680	0.853	0.757
	AD	0.973	0.566	0.564	0.409	0.002	0.025	0.580	0.997	0.733					-				
	FS	0.973	0.999	**0.972**	0.001	0.027	0.000	0.999	0.973	**0.986**	0.973	1.000	**0.973**	0.000	0.027	0.000	1.000	0.973	**0.986**
age	BL	0.973	0.715	0.688	0.285	0.027	0.000	0.707	0.962	0.815	0.855	0.479	0.405	0.450	0.074	0.070	0.474	0.845	0.607
	RB	0.973	0.760	0.734	0.239	0.025	0.002	0.755	0.967	0.848	0.968	0.595	0.569	0.399	0.026	0.006	0.588	0.956	0.728
	FS	0.973	0.983	**0.957**	0.016	0.025	0.002	0.984	0.974	**0.979**	0.973	1.000	**0.973**	0.000	0.027	0.000	1.000	0.973	**0.986**
race	BL	0.973	0.465	0.438	0.535	0.027	0.000	0.450	0.942	0.610	0.855	0.763	0.646	0.209	0.116	0.029	0.756	0.848	0.799
	RB	0.973	0.591	0.566	0.407	0.025	0.002	0.581	0.957	0.723	0.968	0.696	0.684	0.284	0.013	0.020	0.707	0.982	0.822
	FS	0.973	0.995	**0.968**	0.006	0.027	0.000	0.994	0.973	**0.984**	0.973	1.000	**0.973**	0.000	0.027	0.000	1.000	0.973	**0.986**
(age race gender)	BL	0.973	0.668	0.641	0.332	0.027	0.000	0.659	0.960	0.781	0.855	0.455	0.382	0.473	0.073	0.072	0.447	0.840	0.583
	RB	0.973	0.729	0.703	0.270	0.025	0.002	0.723	0.965	0.827	0.968	0.590	0.564	0.404	0.026	0.006	0.583	0.956	0.724
	FS	0.973	1.000	**0.973**	0.000	0.027	0.000	1.000	0.973	**0.986**	0.973	1.000	**0.973**	0.000	0.027	0.000	1.000	0.973	**0.986**

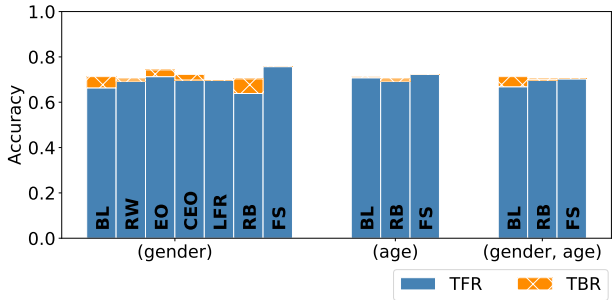
the highest extent without sacrificing their accuracy.

A.2.3. German Credit

It can be seen in Table 8 that our fair Siamese approach achieves the highest fair-F1 scores for the FCNN and LR classifiers on the German Credit dataset under various setting of protected attributes. It improves the individual fairness of the BL models to the highest extent without sacrificing their accuracy. The observations on true bias and false fairness, discussed in Section 5.2, can be made similarly for the German Credit dataset, as shown in Figure 6 and Figure 7.

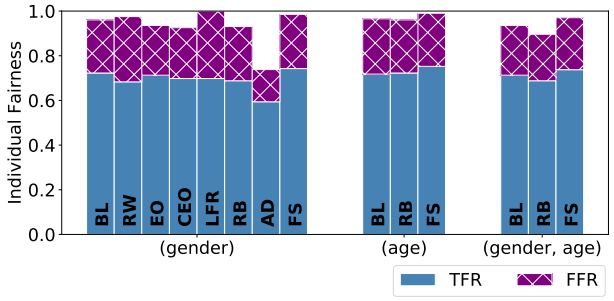


(a) FCNN classifiers for German Credit

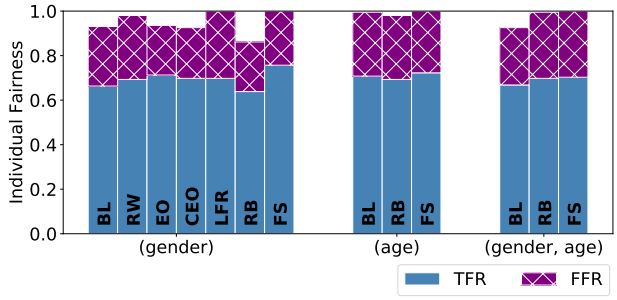


(b) LR classifiers for German Credit

Figure 6. TFR and Accuracy



(a) FCNN classifiers for German Credit



(b) LR classifiers for German Credit

Figure 7. TFR and IF

Table 8. Statistics of FCNN and LR Classifiers for German Credit

Attr	Model	FCNN									LR								
		Acc	IF	TFR	TBR	FFR	FBR	F-P	F-R	F-F1	Acc	IF	TFR	TBR	FFR	FBR	F-P	F-R	F-F1
gender	BL	0.738	0.960	0.723	0.015	0.238	0.025	0.980	0.753	0.851	0.713	0.931	0.663	0.050	0.267	0.020	0.931	0.713	0.807
	RW	0.693	0.975	0.683	0.010	0.292	0.015	0.986	0.701	0.819	0.703	0.980	0.693	0.010	0.287	0.010	0.986	0.707	0.824
	EO	0.743	0.936	0.713	0.030	0.223	0.035	0.960	0.762	0.850	0.743	0.936	0.713	0.030	0.223	0.035	0.960	0.762	0.850
	CEO	0.723	0.926	0.698	0.025	0.228	0.050	0.966	0.754	0.847	0.723	0.926	0.698	0.025	0.228	0.050	0.966	0.754	0.847
	LFR	0.698	1.000	0.698	0.000	0.302	0.000	1.000	0.698	0.822	0.698	1.000	0.698	0.000	0.302	0.000	1.000	0.698	0.822
	RB	0.708	0.931	0.688	0.020	0.243	0.050	0.972	0.739	0.840	0.703	0.861	0.639	0.064	0.223	0.074	0.909	0.741	0.817
	AD	0.713	0.738	0.594	0.119	0.144	0.144	0.833	0.805	0.819					-				
	FS	0.748	0.985	**0.743**	0.005	0.243	0.010	0.993	0.754	**0.857**	0.757	1.000	**0.757**	0.000	0.243	0.000	1.000	0.757	**0.862**
age	BL	0.738	0.965	0.718	0.020	0.248	0.015	0.973	0.744	0.843	0.713	0.995	0.708	0.005	0.287	0.000	0.993	0.711	0.829
	RB	0.738	0.960	0.723	0.015	0.238	0.025	0.980	0.753	0.851	0.703	0.980	0.693	0.010	0.287	0.010	0.986	0.707	0.824
	FS	0.753	0.990	**0.753**	0.000	0.238	0.010	1.000	0.760	**0.864**	0.723	1.000	**0.723**	0.000	0.277	0.000	1.000	0.723	**0.839**
(age gender)	BL	0.738	0.936	0.713	0.025	0.223	0.040	0.966	0.762	0.852	0.713	0.926	0.668	0.045	0.257	0.030	0.938	0.722	0.816
	RB	0.738	0.896	0.688	0.050	0.208	0.055	0.933	0.768	0.842	0.703	0.995	0.698	0.005	0.297	0.000	0.993	0.702	0.822
	FS	0.757	0.970	**0.738**	0.020	0.233	0.010	0.974	0.760	**0.854**	0.703	1.000	**0.703**	0.000	0.297	0.000	1.000	0.703	**0.826**