Accurate Fairness: Improving Individual Fairness without Trading Accuracy

Xuran Li, Peng Wu*, Jing Su

State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China
{lixr, wp, sujing}@ios.ac.cn

Abstract

Accuracy and individual fairness are both crucial for trustworthy machine learning, but these two aspects are often incompatible with each other so that enhancing one aspect may sacrifice the other inevitably with side effects of true bias or false fairness. We propose in this paper a new fairness criterion, *accurate fairness*, to align individual fairness with accuracy. Informally, it requires the treatments of an individual and the individual's similar counterparts to conform to a uniform target, i.e., the ground truth of the individual. We prove that accurate fairness also implies typical group fairness criteria over a union of similar sub-populations. We then present a *Siamese fairness* in-processing approach to minimize the accuracy and fairness losses of a machine learning model under the accurate fairness constraints. To the best of our knowledge, this is the first time that a Siamese approach is adapted for bias mitigation. We also propose fairness confusion matrix-based metrics, *fair-precision*, *fair-recall*, and *fair-F1 score*, to quantify a trade-off between accuracy and individual fairness. Comparative case studies with popular fairness datasets show that our Siamese fairness approach can achieve on average 1.02%-8.78% higher individual fairness (in terms of fairness through awareness) and 8.38%-13.69% higher accuracy, as well as 10.09%-20.57% higher true fair rate, and 5.43%-10.01% higher fair-F1 score, than the state-of-the-art bias mitigation techniques. This demonstrates that our Siamese fairness approach can indeed improve individual fairness without trading accuracy. Finally, the accurate fairness criterion and Siamese fairness approach are applied to mitigate the possible service discrimination with a real Ctrip dataset, by on average fairly serving 112.33% more customers (specifically, 81.29% more customers in an accurately fair way) than baseline models.

Introduction

Machine learning aided intelligent systems have exhibited competitive performances in decision-making tasks such as loan granting (Hardt, Price, and Srebro 2016), criminal justice risk assessment (Berk et al. 2021), and online recommendations (Lambrecht and Tucker 2019). However, the widespread deployments of such machine-learning systems have also spawned social and political concerns, particularly on the fairness of the decisions or predictions made by these systems.

Accuracy and fairness are both crucial for trustworthy machine learning (Huang et al. 2022b,a; Zhang et al. 2021; Su et al. 2022; Makhlouf, Zhioua, and Palamidessi 2021), but these two aspects may be incompatible fundamentally from their own unilateral perspectives, that is, enhancing one aspect may sacrifice the other inevitably with unacceptable consequences (Dutta et al. 2020; Kim, Chen, and Talwalkar 2020; Pinzón et al. 2022). For instance, more accurate predictions on loan applicants' incomes can benefit banks with less lending risks, but the underlying ground truth distribution may tend to prefer applicants with the majority or privileged backgrounds, due to historical practices. Thus, accurate predictions would reflect, even exaggerate such discrimination against minority or unprivileged applicants. In contrast, enhancing just fairness, e.g., by blindly enforcing all the applicants to have the same access to loans, would result in trivially fair but unsound predictions for actually non-qualified applicants. Therefore, accurate but biased, and fair but faulty predictions do not yield a mutually beneficial trade-off between accuracy and fairness. Such incompatibility has recently been shown in (Pinzón et al. 2022) specifically between non-trivial accuracy and equal opportunity, a group fairness criterion.

In this paper, we propose a new fairness criterion, *accurate fairness*, to align individual fairness (Dwork et al. 2012; Galhotra, Brun, and Meliou 2017) with accuracy by uniformly bounding both the accuracy difference and the fairness difference for *similar* sub-populations. Any two individuals are *similar* if both differ only on their sensitive attributes, e.g., genders, races, and ages. Then, an individual is treated by a machine learning model in an *accurately fair* way, if its prediction results for both the individual and the individual's similar counterparts conform to the ground truth of the individual; otherwise, the prediction result for this individual is either faulty or biased. Thus, under the notion of accurate fairness, an individual and the individual's similar counterparts shall be treated similarly in conformance with a uniform target (i.e., the ground truth of the individual), without acknowledging their differences in the sensitive attributes.

As an individual-level fairness criterion, accurate fairness refines the general definition of individual fairness

[*] Corresponding author: Peng Wu

(Dwork et al. 2012) by explicitly focusing on similar sub-populations, where the individuals are exactly the same on their non-sensitive attributes, instead of on any individuals that are close to each other. Thus, fair but faulty predictions can be potentially reduced because a machine learning model does not have to learn "fair" predictions for close individuals, who though differ on some of their non-sensitive attributes, without regard to their different ground truths. Accurate fairness further captures the intuition that fairness criteria shall be truthfully built upon accurate predictions. Consequently, as a by-product, we show that accurate fairness implies group fairness, specifically statistical parity (Calders, Kamiran, and Pechenizkiy 2009; Kamiran and Calders 2009) and confusion matrix-based fairness (Caton and Haas 2020)) over a union of similar sub-populations.

We then present and implement a Siamese fairness approach to mitigate individual bias without trading accuracy. It simultaneously receives multiple similar individuals as training inputs, and aims to minimize the accuracy and fairness losses of a machine learning model (e.g., a neural network model, a logistic regression model, or a support vector machine) under the accurate fairness constraints. We further propose a *fairness confusion matrix* to evaluate how well a machine learning model can balance accuracy with individual fairness, yielding *fair-precision*, *fair-recall*, and *fair-F1 score* metrics. Fair-precision is the proportion of individually fair predictions in accurate predictions, while fair-recall is the proportion of accurate predictions in individually fair predictions. Fair-F1 score is the harmonic mean of fair-precision and fair-recall.

Empirical studies with popular fairness datasets Adult (UCI Machine Learning Repository 1996), German Credit (Hofmann 1994) and ProPublica Recidivism (Angwin et al. 2022), show that the accurate fairness criterion contributes well to delivering a truthfully fair solution for decision-making, and balances accuracy and individual fairness in a win-win manner. Compared with the state-of-the-art bias mitigation techniques, our Siamese fairness approach can on average promote the individual fairness (fairness through awareness) of a machine learning model 1.02%-8.78% higher, and the model accuracy 8.38%-13.69% higher, with 10.09%-20.57% higher true fair rate and 5.43%-10.01% higher F-F1 score.

Finally, we apply the accurate fairness criterion to evaluate a service discrimination problem with a real dataset (Ctrip 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 the same room services, irrespective of their consumption habits. Two neural network models are trained as baseline models, which do suffer service discrimination against customers with different consumption habits. Our Siamese fairness approach can mitigate such discrimination to a great extent, by on average fairly serving 93.00% customers (112.33% more than the baseline models). More importantly, 81.29% more customers are served in an accurately fair way.

The main contributions of this paper are as follows.

- We propose an individual level fairness criterion, *accurate fairness*, such that any individual and the individual's similar counterparts shall all be treated similarly up to the ground truth of the individual. This makes it a new individual fairness criterion that is accuracy-enhanced and can imply certain group-level fairness criteria in the context of sub-populations.
- We present and implement a Siamese fairness approach to optimize the accurate fairness of a machine learning model, by taking similar individuals as parallel training inputs. To the best of our knowledge, this is the first time that a Siamese approach is adapted for individual bias mitigation.
- The accurate fairness criterion and the Siamese fairness approach are applied with popular fairness datasets and a real Ctrip dataset, under the evaluation of what we propose as fairness confusion matrix-based metrics: fair-precision, fair-recall, and fair-F1 score. The case studies reveal the defects of true bias and false fairness in the learned classifiers. Our approach can indeed mitigate these defects and improve individual fairness without trading accuracy.

The rest of this paper is organized as follows. We briefly discuss the related work in Section 2, followed by the formal definition and discussion of the accurate fairness criterion in Section 3. We present the Siamese fairness 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.

Related Work

Fairness Criteria

Fairness criteria presented in the literature are usually partitioned into two categories: group fairness and individual fairness. Please refer to (Galhotra, Brun, and Meliou 2017; Dwork et al. 2012; Kusner et al. 2017; Caton and Haas 2020; Makhlouf, Zhioua, and Palamidessi 2021; Berk et al. 2021) for a comprehensive survey about machine learning fairness notions.

Group fairness criteria concern equal treatments for subgroups with the same sensitive attribute values, and hence are usually defined statistically in terms of conditional independence. Statistical parity (Calders, Kamiran, and Pechenizkiy 2009; Kamiran and Calders 2009, 2012) requires predictions independent of sensitive attributes so that all the sub-groups have the same positive prediction rates. Confusion matrix-based fairness criteria, e.g., equality odds (Hardt, Price, and Srebro 2016) and accuracy equality (Berk et al. 2021), require predictions independent of sensitive attributes under the given ground truths. However, group fairness criteria may be satisfied with carefully selected individuals, who are unfavorably discriminated against in contrast to their similar counterparts (Makhlouf, Zhioua, and Palamidessi 2021). Thus, individual fairness for these individuals is unnecessarily neglected.

Individual fairness criteria can be defined qualitatively or quantitatively by interpreting the notions of *similar individuals* and *similar treatments* (Lahoti, Gummadi, and Weikum

2019b), in order to assess whether similar individuals are treated similarly. Causal discrimination (Galhotra, Brun, and Meliou 2017; Xie and Wu 2020) is such a qualitative definition, where similar individuals are those who differ only on sensitive attributes, and only equal predictions are accounted as similar treatments.In a quantitative or algorithmic definition, task-specific distance metrics are adapted to characterize the similarities between individuals and between prediction outcomes. Fairness through awareness (Dwork et al. 2012) requires that the similarity distance between individuals lays an upper bound on the similarity distance between the corresponding prediction outcomes by the Lipschitz condition.

Individual fairness criteria concern directly the predictions themselves, which nonetheless are possibly (partly) faulty. Accurate fairness presented in this paper refactors the individual fairness criteria from a viewpoint of accuracy to clarify and quantify such incompatibility and also implies certain group fairness criteria over sub-populations.

Bias Mitigation

As summarized in (Caton and Haas 2020; Bellamy et al. 2019), the 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 (Petersen et al. 2021).

Pre-processing methods aim to learn non-discriminative data representations (Sharma et al. 2020). A fair representation learning (LFR) approach (Zemel et al. 2013) obfuscates any information about sensitive attributes in the learned data representation. iFair (Lahoti, Gummadi, and Weikum 2019a) minimizes the data loss to reconcile individual fairness with application utility.

In-processing methods train a machine learning model with fairness as an additional optimization goal. SenSR (Yurochkin, Bower, and Sun 2020) improves the sensitive subspace robustness against certain sensitive perturbations through a distributionally robust optimization approach. SenSeI (Yurochkin and Sun 2021) enforces the treatment invariance on certain sensitive sets by minimizing a transport-based regularizer through a stochastic approximation algorithm.

These methods separate model accuracy from mitigating individual bias and hence may unilaterally improve individual fairness with accuracy decreasing. Our Siamese fairness approach minimizes the model accuracy and fairness losses uniformly subject to the new accurate fairness criterion, thus mitigating individual bias does not necessarily trade accuracy.

Accurate Fairness

We present in this section the notion of accurate fairness and discuss its connections with other individual fairness and group fairness criteria.

Assume a finite and labeled dataset $V \subseteq X \times A$ with the domains of the non-sensitive attributes, the sensitive attributes, and the ground truth labels denoted A, X, Y, respectively. Each input $(x, a) \in V$ is associated with a ground truth label $y \in Y$. Let $I(x, a) \subseteq X \times A$ be the similar sub-population of (x, a), which is the set of the individuals sharing the same non-sensitive attributes values with (x, a), i.e.,

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

Obviously, $(x, a) \in I(x, a)$. Let $card(S)$ be the cardinal number of set S.

Let $f : X \times A \to Y$ denote a classifier learned from a training dataset V, and $\hat{y} = f(x, a)$ the prediction result of classifier f for input (x, a). Then, the accurate fairness criterion can be defined as follow.

Definition 1 (Accurate Fairness)**.** A classifier $f : X \times A \to Y$ is *accurately fair* to an input $(x, a) \in V$, if for any individual $(x, a') \in I(x, a)$, the distance $D(y, f(x, a'))$ between the ground truth y of input (x, a) and the prediction result $f(x, a')$ is at most $K \geq 0$ times of the distance $d((x, a), (x, a'))$ between (x, a) and (x, a'), i.e.,

$$D(y, f(x, a')) \leq Kd((x, a), (x, a')) \tag{1}$$

where $D(\cdot, \cdot)$ and $d(\cdot, \cdot)$ are distance metrics.

Herein, the accurate fairness constraint (1) captures uniformly the accuracy and individual fairness requirements with respect to input $(x, a) \in V$:

- (Accuracy) Since $(x, a) \in I(x, a)$, constraint (1) reduces to $D(y, f(x, a)) = 0$ (i.e., $y = f(x, a)$ due to the identity of indiscernibles of a distance metric) for input (x, a) itself;
- (Individual Fairness) For any similar individual $(x, a') \in I(x, a)$ with $a' \neq a$, constraint (1) reduces the Lipschitz condition (Dwork et al. 2012) $D(f(x, a), f(x, a')) \leq Kd((x, a), (x, a'))$ for similar individual (x, a') within sub-population $I(x, a)$, as shown by the following theorem.

Theorem 1. *If a classifier $f : X \times A \to Y$ is accurately fair to an input $(x, a) \in V$, then*

$$D(f(x, a), f(x, a')) \leq Kd((x, a), (x, a'))$$

for any similar individual $(x, a') \in I(x, a)$ with $a' \neq a$.

Proof. Due to the triangle inequality and symmetry of a distance metric,

$$D(f(x, a), f(x, a')) \leq K(D(y, f(x, a)) + D(y, f(x, a')))$$

where $a' \neq a$ and y is the ground truth of input $(x, a) \in V$.

Then, By Definition 1, if classifier f is accurately fair to input (x, a), for any similar individual $(x, a') \in I(x, a)$ with $a' \neq a$,

$$D(y, f(x, a)) \leq Kd((x, a), (x, a)) = 0$$

$$D(y, f(x, a')) \leq Kd((x, a), (x, a'))$$

Thus, $D(f(x, a), f(x, a')) \leq Kd((x, a), (x, a'))$ □

It can be seen from Theorem 1 that accurate fairness refactors the general definition of individual fairness (Dwork et al. 2012) over similar sub-populations on the basis of accuracy.

Accurate fairness also collectively endorses group fairness criteria over the union of similar sub-populations. Consider the following definition of accurate parity, which is a qualitative version of accurate fairness.

Definition 2 (Accurate Parity)**.** A classifier $f : X \times A \to Y$ is *accurately equal* to an input $(x, a) \in V$, if for any individual $(x, a') \in I(x, a)$,

$$y = f(x, a) = f(x, a') \tag{2}$$

where y is the ground truth of input(x, a).

Obviously accurate parity entails accurate fairness because with the accurate parity constraint (2), $D(y, f(x, a')) = 0$ for any individual $(x, a') \in I(x, a)$.

Let $\mathbf{X}$, $\mathbf{A}$, $\mathbf{Y}$, and $\hat{\mathbf{Y}}$ denote the random variables representing the non-protected attributes, the protected attributes, the ground truths, and the prediction results. The following theorem shows that accurate parity implies statistical parity and confusion matrix-based fairness over $I(W) = \cup_{(x,a)\in W} I(x, a)$ for certain $W \subseteq V$. As far as the group fairness criteria are concerned, assume each individual $(x, a') \in I(W)$ is associated with the same ground truth as some $(x, a) \in W$ is.

Theorem 2. *If a classifier $f : X \times A \to Y$ is accurately equal to each input in $W \subseteq V$, then f satisfies statistical parity and confusion matrix-based fairness over $I(W)$.*

Proof. The accurate parity constraint (2) implies that

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

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

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

over $I(W)$. Thus, the accuracy (or inaccuracy) for the similar individuals are independent on the sensitive attributes. Therefore, f satisfies the confusion matrix-based fairness over $I(W)$. Furthermore, f also satisfies statistical parity over $I(W)$, because $\mathbf{P}(\hat{\mathbf{Y}} = \hat{y}|\mathbf{A} = a) = \mathbf{P}(\hat{\mathbf{Y}} = \hat{y}, \hat{\mathbf{Y}} = \mathbf{Y}|\mathbf{A} = a) + \mathbf{P}(\hat{\mathbf{Y}} = \hat{y}, \hat{\mathbf{Y}} \neq \mathbf{Y}|\mathbf{A} = a)$ for any $\hat{y} \in Y$. □

Note that Proposition 3 in (Barocas, Hardt, and Narayanan 2019) shows that statistical parity (independence) and confusion matrix-based fairness (separation) cannot both hold unless $\mathbf{A} \perp \mathbf{Y}$ or $\hat{\mathbf{Y}} \perp \mathbf{Y}$, while the former is admitted on $I(W)$ under the accurate parity constraints.

Generally speaking, accurate fairness lays an upper bound on the treatment differences between groups with different sensitive attribute values. Let $I_a(W)$ be the set of individuals in $I(W)$ with the same sensitive attribute values $a \in A$, i.e., $I_a(W) = \{(x, a)|(x, a) \in I(W)\}$.

Theorem 3. *If a classifier $f : X \times A \to Y$ is accurately fair to each input in $W \subseteq V$, for any $(x, a^*) \in W$, $a \in A$ and $a \neq a^*$, then over $I_a(W)$,*

$$\begin{array}{rcl} \mathbf{E}[D(\mathbf{Y}, f(\mathbf{X}, a))] & \leq & K\mathbf{E}[d((\mathbf{X}, a^*), (\mathbf{X}, a))] \\ \mathbf{E}[D(f(\mathbf{X}, a^*), f(\mathbf{X}, a))] & \leq & K\mathbf{E}[d((\mathbf{X}, a^*), (\mathbf{X}, a))] \end{array}$$

Proof. Recall that over $I_a(W)$, $\mathbf{E}[D(\mathbf{Y}, f(\mathbf{X}, a))] = \sum_{(x,a)\in I_a(W)} \mathbf{P}(y, x, a) D(y, f(x, a))$, $\mathbf{E}[d((\mathbf{X}, a^*), (\mathbf{X}, a))] = \sum_{(x,a)\in I_a(W)} \mathbf{P}(x, a) d((x, a^*), (x, a))$, and $\mathbf{E}[D(f(\mathbf{X}, a^*), f(\mathbf{X}, a))] = \sum_{(x,a)\in I_a(W)} \mathbf{P}(x, a) D(f(x, a^*), f(x, a))$, where y is the ground truth of (x, a^*), $\mathbf{P}(y, x, a)$ is the joint probability of $\mathbf{Y} = y$ and $(\mathbf{X}, \mathbf{A}) = (x, a)$, and $\mathbf{P}(x, a)$ is the probability of $(\mathbf{X}, \mathbf{A}) = (x, a)$.

Then, since f is accurately fair to each input in W, the proof is concluded by Definition 1. □

Theorems 3 suggest that under the accurate fairness criterion, the treatment differences between individuals and their similar counterparts are also bounded by the distances between these individuals themselves, hence leading to quantitatively fair treatments over groups with different sensitive attributes values.

Siamese In-Processing for Accurate Fairness

We present in this section the Siamese fairness in-processing approach to achieve accurate fairness. It intends to train a machine learning model for the following optimization problem, which minimizes the cumulative loss for the union $I(V)$ of similar populations, subject to the accurate fairness constraints.

$$\begin{array}{c} \min\limits_{f} \sum\limits_{(x_i,a_i)\in V} \sum\limits_{(x_i,a_{ij})\in I(x_i,a_i)} L(y_i, f(x_i, a_{ij})) \\ \text{s.t. } D(y_i, f(x_i, a_{ij})) \leq Kd((x_i, a_i), (x_i, a_{ij})) \\ \text{for any } (x_i, a_i) \in V, (x_i, a_{ij}) \in I(x_i, a_i) \end{array} \tag{3}$$

where y_i is the ground truth of $(x_i, a_i) \in V$, $1 \leq i \leq card(V)$, $1 \leq j \leq card(I(x_i, a_i))$ and $L(\cdot, \cdot)$ is a loss function for training the machine learning model.

By appealing to Karush–Kuhn–Tucker conditions (Boyd and Vandenberghe 2004), it is equivalent to solve the following max-min optimization problem with the Lagrange multipliers $\lambda_{ij} \geq 0$ for each $(x_i, a_i) \in V, (x_i, a_{ij}) \in I(x_i, a_i)$, assuming that the loss function $L(\cdot, \cdot)$ and the distance metrics $D(\cdot, \cdot)$ and $d(\cdot, \cdot)$ are all convex.

$$\begin{array}{l} \max\limits_{\lambda} \min\limits_{f} \sum\limits_{(x_i,a_i)\in V} \sum\limits_{(x_i,a_{ij})\in I(x_i,a_i)} \Big(L(y_i, f(x_i, a_{ij})) + \\ \qquad \lambda_{ij} \big(D(y_i, f(x_i, a_{ij})) - Kd((x_i, a_i), (x_i, a_{ij})) \big) \Big) \end{array} \tag{4}$$

It can be seen that the objective function in (4) renders a possibility of stochastic estimation with observations on the union $I(V)$ of the similar sub-populations, instead of just the training dataset $V \subseteq I(V)$. A Siamese network can accept multiple inputs in parallel to train multiple models with shared parameters (Chopra, Hadsell, and LeCun 2005). Thus, it provides a training mechanism to treat individuals in a similar sub-population in a uniform manner.

Therefore, we propose to adapt a Siamese network for accurate fairness in-processing. The architecture of our approach is shown in Figure 1. It first generates the similar sub-population $I(x_i, a_i)$ for each training input $(x_i, a_i) \in V$ with $1 \leq i \leq card(V)$ through fair augmentation. Then, it trains $m = card(I(x_i, a_i))$ copies of a machine learning model with shared parameters θ for the sake of minimizing

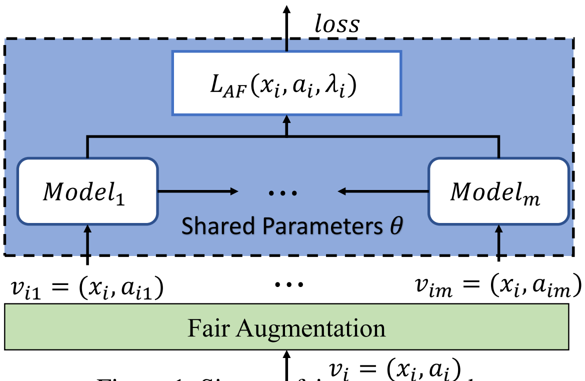

Figure 1: Siamese fairness approach

the accurate fairness loss $L_{AF}(x_i, a_i, \lambda_i)$ over the similar sub-population $I(x_i, a_i)$:

$$L_{AF}(x_i, a_i, \lambda_i) = \sum_{(x_i, a_{ij}) \in I(x_i, a_i)} \Big(L(y_i, f(x_i, a_{ij})) + \lambda_{ij}(D(y_i, f(x_i, a_{ij})) - Kd((x_i, a_i), (x_i, a_{ij}))) \Big)$$

where $\lambda_i = (\lambda_{i1}, \cdots, \lambda_{ij}, \cdots, \lambda_{im})$.

Algorithm 1 shows the workflow of our Siamese fairness approach in detail. At Lines 1-6, the training dataset V is augmented with the similar counterparts of each input $v_i = (x_i, a_i) \in V$, resulting in $I(V)$ for the subsequent Siamese training. At Lines 10-12, each $Model_j$ in Figure 1 run a copy of classifier f_θ with the shared parameters θ, accepting the j-th similar individual $(x_i, a_{ij}) \in I(x_i, a_i)$ and producing $f_\theta(x_i, a_{ij})$ for $1 \le j \le card(I(x_i, a_i))$. At Lines 13-16, the Lagrange multipliers λ_i and the shared parameters θ for v_i are obtained by applying an error Back-Propagation (BP) algorithm (Werbos 1974; Rummelhart, Hinton, and Williams 1986a,b) to optimize $\sum_{(x_i,a_i)\in V} L_{AF}(x_i, a_i, \lambda_i)$ (i.e., the objective function in (4)).

The Siamese in-processing architecture in Figure 1 allows treating individuals in one similar sub-population simultaneously and uniformly as a whole during each iteration of back-propagation in Algorithm 1, while classical training algorithms usually handle inputs one by one, unable to accommodate the accurate fairness criterion for bias mitigation.

Implementation and Evaluation

We implement the Siamese fairness approach (Algorithm 1) in Python 3.8 with TensorFlow 2.4.1. Our 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, in comparison with the state-of-the-art individual fairness bias mitigation techniques with regard to binary or multi-valued sensitive attributes, or the combinations thereof. The source code and the experimental datasets and models are available at https://github.com/Xuran-LI/AccurateFairnessCriterion.

Datasets and Models

The three popular fairness datasets Adult, German Credit and COMPAS, and a real dataset from Ctrip are used for the evaluation. The instances with unknown or empty values have been removed from the datasets before training. Table 1 reports the size and the sensitive attributes of each dataset, and the models trained with these datasets. “$A(m)$” means that attribute A has m values. An FCNN(l) model is a fully connected neural network (FCNN) classifier with l layers; while an LR or SVM model is a logistic regression (LR) or Support Vector Machine (SVM) classifier, respectively. These classifiers are referred to as the baseline (BL) models in the evaluation.

Algorithm 1: Siamese Fairness (SF)

```
Input: dataset $V$, classifier $f_\theta$, learning rate $\eta$
Output: classifier $f_\theta$
 1: for each $v_i = (x_i, a_i) \in V$ do
 2:     $I(x_i, a_i) \leftarrow \{v_i\}$;
 3:     for each $a \in A$ and $a \neq a_i$ do
 4:         $I(x_i, a_i) \leftarrow I(x_i, a_i) \cup \{(x_i, a)\}$
 5:     end for
 6: end for
 7: Initialize $\lambda_1, \cdots, \lambda_{card(V)}$ and $\theta$;
 8: repeat
 9:     for each $v_i = (x_i, a_i) \in V$ do
10:         for each $(x_i, a_{ij}) \in I(x_i, a_i)$ do
11:             Compute $f_\theta(x_i, a_{ij})$;
12:         end for
13:         for each $\lambda_{ij} \in \lambda_i, w \in \theta$ do
14:             $\lambda_{ij} \leftarrow \max(0, \lambda_{ij} + \eta \frac{\partial L_{AF}(x_i,a_i,\lambda_i)}{\partial \lambda_{ij}})$;
15:             $w \leftarrow w - \eta \frac{\partial L_{AF}(x_i,a_i,\lambda_i)}{\partial w}$
16:         end for
17:     end for
18: until $\theta$ converges or the maximal number of iterations
    is reached
19: return $f_\theta$
```

Evaluation Metrics

In addition to the accuracy metric (ACC), the group fairness metrics (including statistical parity difference (SPD),

Dataset	Size	Models	Sensitive Attributes

Adult
(Census
Income)
	45222	FCNN(3)
		LR
		SVM
			gender(2)
			age(71)
			race(5)

German
Credit
	1000	FCNN(3)
		LR
		SVM
			gender(2)
			age(51)

ProPublica
Recidivism
(COMPAS)
	6172	FCNN(3)
		LR
		SVM
			gender(2)
			age(71)
			race(6)

Ctrip	68191	FCNN(3)
		FCNN(5)
			6 customer
			consumption habits

Table 1: Datasets and Models

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

Table 2: Fairness Confusion Matrix

equal odds difference (EOD), and average odds difference (AVOD)), and the individual fairness metrics (including fairness through awareness (FTA), consistency (CON)), we propose a fairness confusion matrix (as shown in Table 2), and the following fairness confusion matrix based metrics to evaluate the bias mitigation performance of a machine learning model in balancing its accuracy with individual fairness.

Definition 3 (Fairness Confusion Matrix Based Metrics)**.** For classifier $f : X \times A \to Y$ and input $(x, a) \in V$,

- the prediction $f(x, a)$ is *true fair* if the prediction $f(x, a')$ for any $(x, a') \in I(x, a)$ conforms to y, the ground truth of (x, a);
- $f(x, a)$ is *true biased* if it conforms to y, but the prediction $f(x, a')$ for some $(x, a') \in I(x, a)$ with $a' \neq a$ does not;
- $f(x, a)$ is *false fair* if it does not conform to y, but is consistent to the prediction $f(x, a')$ for any $(x, a') \in I(x, a)$ with $a' \neq a$;
- $f(x, a)$ is *false biased* if neither the predictions $f(x, a')$ for all $(x, a') \in I(x, a)$ are consistent to each other, nor $f(x, a)$ conforms to y.

Let *True Fair Rate* (TFR), *True Biased Rate* (TBR), *False Fair Rate* (FFR), *False Biased Rate* (FBR) be the proportion of the true fair, true biased, false fair, false biased predictions in all the predictions on V, respectively. Then, *Fair-Precision* (F-P), *Fair-Recall* (F-R) and *Fair-F1 Score* (F-F1) can be defined as follows:

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

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

Informally, the fairness confusion matrix summarizes the orthogonal synergy between individual fairness and accuracy. $F\text{-}P$ measures the individually fair proportion in the accurate predictions, while $F\text{-}R$ measures the accurate proportion in the individually fair predictions. $F\text{-}F1$ combines fair-precision and fair-recall to measure the compatibility between accuracy and individual fairness.

Mitigating Individual Bias

Table 3 reports the average statistics of ten runs for each bias mitigation approach compared over the three fairness datasets. Columns iFair, LFR, SSI, SSR, SF, and SF_3 show the performances of the FCNN classifiers by applying iFair (Lahoti, Gummadi, and Weikum 2019a), LFR (Bellamy et al. 2019; Zemel et al. 2013), SenSeI (Yurochkin and Sun 2021), SenSR (Yurochkin, Bower, and Sun 2020), and our Siamese fairness approach on the baseline models, respectively. For the SF models, all the sensitive attribute values are used (whenever applicable) for augmentation, while for the SF_3 models, only the maximum and minimum values of the sensitive attributes are used for augmentation to save computation consumption. The metrics with subscript $I(V)$ are computed over an augmented dataset $I(V)$, instead of an original (default) dataset V. For our SF and SF_3 models, we use the Mean Squared Error loss function for the FCNN and LR classifiers, and the Hinge loss function for the SVM ones. The distance metrics in the accurate fairness constraints are all implemented with the Mean Absolute Error. An Adam optimizer (Kingma and Ba 2015) is deployed for training the FCNN and SVM classifiers, while a gradient descent optimizer (Sutskever et al. 2013; Robbins and Monro 1985) is for training the LR ones.

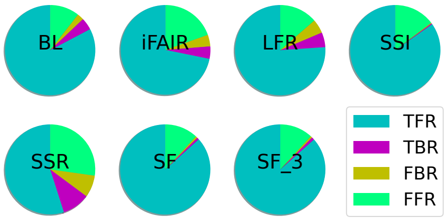

Figure 2: Fairness Confusion Matrix performances (FCNNs)

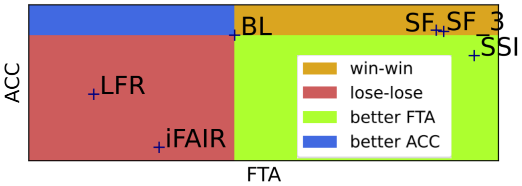

Figure 3: Fairea evaluation (FCNNs)

Figure 2 and Figure 3 further demonstrates the fairness-accuracy trade-offs in terms of fairness confusion matrix performances and the Fairea evaluation (Hort et al. 2021), respectively.

It can be seen in Table 3 and Figure 2 that our Siamese fairness approach achieves the highest TFR and F-F1 performances, with both accuracy (ACC) and individual fairness (FTA) well improved. Compared with the state-of-the-art individual bias mitigation approaches, our Siamese fairness approach on average promotes 0.105 ACC (13.69% higher ACC) and 0.080 FTA (8.78% higher FTA) of a classifier. This is a direct consequence of the observation that our Siamese fairness approach promotes accurate fairness with on average 0.148 TFR (20.57% higher TFR), reducing 0.043 TBR (85.03% lower TBR), 0.037 FBR (85.07% lower FBR), 0.069 FFR (36.53% lower FFR), and promoting 0.085 F-F1 score (10.01% higher F-F1 score).

The Fairea evaluation approach also certifies that only the SF and SF_3 models fall into the win-win trade-off region, as shown in Figure 3, which supplies a sufficiently strong signal on the bias mitigation effectiveness of our Siamese fairness approach. The other bias mitigation approaches may improve the individual fairness of a classifier but at the cost of its accuracy. Moreover, our Siamese fairness approach also achieves the highest group fairness over the union of similar sub-populations. This suggests that accurate fairness can help reach a synergy between individual fairness and

Metrics	BL	iFAIR	LFR	SSI	SSR	SF	SF_3
ACC	0.871±0.007	0.763±0.047	0.814±0.020	0.851±0.017	0.648±0.076	0.874±0.004	0.876±0.004
SPD	0.105±0.015	0.110±0.095	0.096±0.059	0.058±0.010	0.077±0.085	0.111±0.015	——
EOD	0.046±0.011	0.109±0.092	0.124±0.054	0.038±0.010	0.086±0.089	0.051±0.016	
AVOD	0.037±0.020	0.093±0.123	0.146±0.076	0.041±0.013	0.088±0.112	0.030±0.017	
$\text{SPD}_{I(V)}$	0.045±0.018	0.079±0.073	0.102±0.073	0.007±0.004	0.080±0.090	0.003±0.002	
$\text{EOD}_{I(V)}$	0.059±0.021	0.086±0.079	0.124±0.082	0.009±0.006	0.083±0.093	0.007±0.004	
$\text{AVOD}_{I(V)}$	0.081±0.028	0.067±0.097	0.149±0.101	0.011±0.006	0.087±0.111	0.005±0.003	
CON	0.928±0.005	0.976±0.018	0.970±0.005	0.957±0.007	0.957±0.041	0.934±0.008	0.935±0.009
FTA	0.933±0.025	0.913±0.075	0.897±0.073	0.994±0.007	0.822±0.189	0.986±0.009	0.984±0.009
TFR	0.827±0.020	0.718±0.075	0.761±0.053	0.848±0.017	0.549±0.151	0.867±0.006	0.866±0.007
TBR	0.044±0.019	0.045±0.047	0.053±0.042	0.003±0.004	0.099±0.109	0.008±0.005	0.010±0.006
FFR	0.105±0.010	0.195±0.056	0.136±0.029	0.146±0.017	0.273±0.090	0.119±0.006	0.118±0.006
FBR	0.024±0.011	0.042±0.040	0.050±0.033	0.003±0.004	0.079±0.085	0.007±0.005	0.006±0.004
F-R	0.884±0.008	0.789±0.055	0.851±0.027	0.853±0.017	0.663±0.104	0.878±0.005	0.879±0.005
F-P	0.947±0.023	0.941±0.062	0.933±0.052	0.996±0.005	0.838±0.187	0.991±0.006	0.988±0.007
F-F1	0.913±0.009	0.854±0.043	0.885±0.022	0.917±0.010	0.725±0.129	0.930±0.003	0.929±0.004

Table 3: Statistics of FCNN classifiers on the three fairness datasets

group fairness, such that improving group fairness can be manifested by improving individual fairness.

Due to the page limit, we herein discuss the performances of the FCNN classifiers. Similar observations can be made on the LR and SVM classifiers. Please refer to (Li, Wu, and Su 2022) for the detailed experimental results.

Service Discrimination with the Ctrip Dataset

We then apply the accurate fairness criterion and the Siamese fairness approach to investigate a service discrimination problem, where customers with different consumption habits may be recommended disparate services, even though they pay the same prices for the same rooms. The Ctrip dataset includes 6 consumption habits attributes of customers (including the average time of order confirmation, the average advance days of booking, the average star level, class level, recommended level of hotels booked, and the average days of hotel stay) and 6 attributes of hotels (including order date, hotel ID, room type, room ID, star level and room price). For the service discrimination problem, the 6 customer attributes are designated as the sensitive attributes. The ground truth labels represent the room service types.

As reported in Table 4, the baseline (BL) models get an accuracy of 66.48% on average, but only 34.68% (TFR) customers are treated both accurately and fairly. Through Siamese fairness in-processing, the average TFR is extremely improved to 62.85%, very close to its upper bound, which is the average accuracy of 65.77%. Our Siamese fairness approach can make most (on average 93.00%) of the customers to be fairly served irrespective of their consumption habits. Thus, for a further truthful promotion of their individual fairness, it is left to improve the accuracy of the classifiers themselves, instead of trading it.

Metric	BL(3)	SF_3(3)	BL(5)	SF_3(5)
ACC	0.664	0.660	0.666	0.655
CON	0.940	0.978	0.927	0.988
FTA	0.524	0.902	0.352	0.958
TFR	0.412	0.620	0.281	0.637
TBR	0.251	0.040	0.385	0.018
FFR	0.112	0.282	0.071	0.321
FBR	0.225	0.058	0.263	0.024
F-R	0.792	0.688	0.804	0.666
F-P	0.621	0.940	0.423	0.972
F-F1	0.689	0.794	0.516	0.790

Table 4: Statistics of FCNNs on the Ctrip datatset

Conclusion

We present in this paper the accurate fairness criterion, based on the intuition that similar sub-populations shall be treated similarly up to the ground truths. It enhances individual fairness from the perspective of accuracy and paves a way to achieve harmony among accuracy, individual fairness, and group fairness. Accurate fairness also induces a fairness confusion matrix that can expose the side effects of trading accuracy for individual fairness and vice versa, i.e., resulting in individually fair but faulty predictions (false fairness), or accurate but individually biased predictions (true bias).

Then we present and evaluate our Siamese fairness in-processing approach (SF) in terms of fairness confusion matrix metrics. It aims to maximize the accurate fairness of a decision-making model with similar sub-populations as parallel training inputs. In this way, SF can significantly improve individual fairness without trading accuracy.

We envisage that the state-of-the-art bias mitigation techniques can be further refined from the perspective of accurate fairness. Studies on individual fairness usually rely on pre-specified sensitive attributes and disadvantaged groups. As part of future work, the fairness confusion matrix can be adapted to analyze which sensitive attributes pose more impacts on prediction outcomes. The accurate fairness criterion can be further utilized to help diagnose which groups under these attributes are treated unfavorably.

Acknowledgments

We thank Mr. Han Yue as well as the anonymous referees for their valuable and insightful comments. This work is partially supported by the National Key R&D Program of China under the Grant No. 2022YFA1005100, 2022YFA1005101, and 2022YFA1005104, the CAS-INRIA Joint Research Program under the Grant No. GJHZ1844, the CAS Project for Young Scientists in Basic Research under the Grant No.YSBR-040, Beijing Academy of Artificial Intelligence (BAAI), and Joint Laboratory on Cyberspace Security, China Southern Power Grid.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2022. Machine Bias. *Ethics of Data and Analytics: Concepts and Cases*, 254.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org.

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. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, 4:1–4:15.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 3–44.

Boyd, S.; and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press. ISBN 0521833787.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building Classifiers with Independency Constraints. In *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops*, 13–18.

Caton, S.; and Haas, C. 2020. Fairness in Machine Learning: A Survey. *CoRR*, abs/2010.04053.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 539–546.

Ctrip. 2019. Ctrip Customer Booking Dataset. https://www.heywhale.com/mw/project/5ca2d6098408c1002b48bf3c/dataset. Accessed: 2022-03-29.

Dutta, S.; Wei, D.; Yueksel, H.; Chen, P.; Liu, S.; and Varshney, K. R. 2020. 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*, 2803–2813.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012*, 214–226.

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

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, 3315–3323.

Hofmann, H. 1994. Statlog (German Credit Data). DOI: 10.24432/C5NC77. Accessed: 2022-03-29.

Hort, M.; Zhang, J. M.; Sarro, F.; and Harman, M. 2021. Fairea: a model behaviour mutation approach to benchmarking bias mitigation methods. In *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 994–1006.

Huang, P.; Yang, Y.; Jia, F.; Liu, M.; Ma, F.; and Zhang, J. 2022a. Word Level Robustness Enhancement: Fight Perturbation with Perturbation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event*, 10785–10793.

Huang, P.; Yang, Y.; Liu, M.; Jia, F.; Ma, F.; and Zhang, J. 2022b. ϵ-weakened robustness of deep neural networks. In *ISSTA '22: 31st ACM SIGSOFT International Symposium on Software Testing and Analysis, Virtual Event*, 126–138.

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

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

Kim, J. S.; Chen, J.; and Talwalkar, A. 2020. FACT: A Diagnostic for Group Fairness Trade-offs. In *Proceedings of the 37th International Conference on Machine Learning*, 5264–5274.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 4066–4076.

Lahoti, P.; Gummadi, K. P.; and Weikum, G. 2019a. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In *35th IEEE International Conference on Data Engineering*, 1334–1345.

Lahoti, P.; Gummadi, K. P.; and Weikum, G. 2019b. Operationalizing Individual Fairness with Pairwise Fair Representations. *Proc. VLDB Endow.*, 506–518.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, (7): 2966–2981.

Li, X.; Wu, P.; and Su, J. 2022. Accurate Fairness: Improving Individual Fairness without Trading Accuracy. *CoRR*, abs/2205.08704.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Inf. Process. Manag.*, 58(5): 102642.

Petersen, F.; Mukherjee, D.; Sun, Y.; and Yurochkin, M. 2021. Post-processing for Individual Fairness. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, 25944–25955.

Pinzón, C.; Palamidessi, C.; Piantanida, P.; and Valencia, F. 2022. On the Impossibility of Non-trivial Accuracy in Presence of Fairness Constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 7993–8000.

Robbins, H.; and Monro, S. 1985. *A Stochastic Approximation Method*. Herbert Robbins Selected Papers.

Rummelhart, D.; Hinton, G.; and Williams, R. 1986a. Learning Internal Representations by Error Propagation. *Nature*, 318–362.

Rummelhart, D.; Hinton, G. E.; and Williams, R. J. 1986b. Learning Representations by Back Propagating Errors. *Nature*, 533–536.

Sharma, S.; Zhang, Y.; Aliaga, J. M. R.; Bouneffouf, D.; Muthusamy, V.; and Varshney, K. R. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society*, 358–364.

Su, J.; Zhang, Z.; Wu, P.; Li, X.; and Zhang, J. 2022. Adversarial Input Detection Based on Critical Transformation Robustness. In *33rd IEEE International Symposium on Software Reliability Engineering*.

Sutskever, I.; Martens, J.; Dahl, G. E.; and Hinton, G. E. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, 1139–1147.

UCI Machine Learning Repository. 1996. Adult. DOI: 10.24432/C5XW20. Accessed: 2022-03-29.

Werbos, P. J. 1974. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science. Thesis (Ph.D.). Appl. Math. Harvard University. *Ph.D. thesis, Harvard University*.

Xie, W.; and Wu, P. 2020. 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)*, 121–128.

Yurochkin, M.; Bower, A.; and Sun, Y. 2020. Training individually fair ML models with sensitive subspace robustness. In *8th International Conference on Learning Representations*.

Yurochkin, M.; and Sun, Y. 2021. SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness. In *9th International Conference on Learning Representations*.

Zemel, R. S.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, 325–333.

Zhang, Z.; Wu, P.; Chen, Y.; and Su, J. 2021. Out-of-Distribution Detection through Relative Activation-Deactivation Abstractions. In *32nd IEEE International Symposium on Software Reliability Engineering*, 150–161.