11-2022

Adaptive fairness improvement based causality analysis

Mengdi ZHANG

Jun SUN
Singapore Management University, junsun@smu.edu.sg

Adaptive Fairness Improvement Based on Causality Analysis

Mengdi Zhang
mdzhang.2019@phdcs.smu.edu.sg
Singapore Management University
Singapore

Jun Sun
junsun@smu.edu.sg
Singapore Management University
Singapore

ABSTRACT

Given a discriminating neural network, the problem of fairness improvement is to systematically reduce discrimination without significantly scarifies its performance (i.e., accuracy). Multiple categories of fairness improving methods have been proposed for neural networks, including pre-processing, in-processing and post-processing. Our empirical study however shows that these methods are not always effective (e.g., they may improve fairness by paying the price of huge accuracy drop) or even not helpful (e.g., they may even worsen both fairness and accuracy). In this work, we propose an approach which adaptively chooses the fairness improving method based on causality analysis. That is, we choose the method based on how the neurons and attributes responsible for unfairness are distributed among the input attributes and the hidden neurons. Our experimental evaluation shows that our approach is effective (i.e., always identify the best fairness improving method) and efficient (i.e., with an average time overhead of 5 minutes).

CCS CONCEPTS

• **Software and its engineering → Extra-functional properties**.

KEYWORDS

Fairness, Machine Learning, Fairness Improvement, Causality Analysis

1 INTRODUCTION

Neural networks have found their way into a variety of systems, including many which potentially have significant societal impact, such as personal credit rating [18], criminal sentencing [6, 40], face recognition [41] and resume shortlisting [39]. While these neural networks often have high accuracy in these classification tasks, some concerning fairness issues have been observed as well [8, 9,

13, 20, 40]. That is, the predictions made by these neural networks may be biased with regard to certain protected attributes such as sex, race, and gender. For instance, it has been shown [7] that a neural network trained to predict people's income level based on an individual's personal information (which can be used in applications such as bank loan approval) is much more likely to predict male individuals with high-income level. Further analysis shows that for many individuals, changing only the gender or race causes the output of the predictions to flip [21]. For another instance, it has been shown [6] that a machine learning model used to predict the recidivism risk for suspected criminals is more likely to mislabel black defendants as having a high recidivism risk.

In recent years, many methods and tools have been proposed to detect discrimination in neural networks systematically (e.g., through the so-called fairness testing [5, 21, 34, 44, 50]), and more relevantly, to improve the fairness of neural networks [3, 12, 19, 23, 28–30, 37, 47]. In general, existing fairness improving methods can be classified into three categories according to when the method is applied, e.g., pre-processing, in-processing and post-processing methods. Pre-processing methods [11, 19, 28, 46] aim to reduce the bias in the training data so as to reduce the bias of model predictions; in-processing methods [3, 4, 12, 30, 31, 47] focus on the model and the training process; and post-processing methods [23, 29, 37] modify the prediction results directly rather than the training data or the model.

However, fairness improving is a complicated task and it is not always clear which method should be applied. As shown in Section 3, different fairness improving methods perform significantly differently on different models (which is consistent with the partial results reported in [8, 13, 20]). More importantly, applying the 'wrong' method would not only lead to a huge loss in accuracy (e.g., the accuracy of the model trained on the COMPAS dataset drops by 35% after applying the Reject Option post-processing method), but also lead to worsened fairness. For instance, out of 90 cases (i.e., combinations of model, protected attribute and fairness improving method) that we examined in Section 3, 20% of them result in worsened fairness. Furthermore, a fairness improving method may be effective with respect to one protected attribute whilst being harmful with respect to another protected attribute. For instance, the fairness of the model trained on Adult Income dataset improves by around 4% with respect to the *gender* attribute after applying the Equalized Odds post-processing method and worsens by 20% with respect to the *race* attribute. Given that many of the fairness improving methods require significant effort and computing resource, it is infeasible to try all of them and identify the best performing one. It is thus important to have a systematic way of identifying the 'right' method efficiently.

In this work, we propose to choose the 'right' fairness improving method based on causality analysis. Intuitively, the idea is to

conduct causality analysis so as to understand the causes of the discrimination, i.e., whether a certain number of input attributes or hidden neurons are highly responsible for the unfairness. Formally, we use the probability of high causal effects and Coefficient of Variation to characterize the distribution of the causal effects. Depending on the result of the causality analysis, we then choose the fairness improving method accordingly. For instance, if a small number of input attributes bare most of the responsibility for unfairness, a pre-processing method such as [19, 28] would be the choice, whereas an in-processing method would be the choice if some neurons are highly responsible. Our approach is designed based on the results of an empirical study which evaluates 9 fairness improving methods (i.e., 2 pre-processing methods, 4 in-processing methods and 3 post-processing methods) on 4 different benchmark datasets with respect to different fairness metrics. Our approach is systematically evaluated with the same models. The results show that our selected processing approach is the optimal choice to improve group fairness in all cases and the optimal choice to reduce individual discrimination in most cases.

The remainders of the paper are organized as follows. In Section 2, we review relevant background. In Section 3, we present results from our empirical study which motivates our approach. In Section 4, we present our adaptive fairness improving method. In Section 5, we evaluate our approach. Lastly, we review related work in Section 7 and conclude in Section 8.

2 BACKGROUND

In the following, we review relevant background on fairness and existing fairness improving methods.

2.1 Fairness Definitions

In the literature, there are multiple definitions of fairness [10, 17, 21, 27, 33, 45]. What is common across different definitions is that to define fairness, we must first identify a set of protected attributes (a.k.a. sensitive attributes). Commonly recognized protected attributes instance race, sex, age and religion. Note that different models may have different protected attributes.

In the following, we introduce two popular definitions of fairness, i.e., group fairness and individual discrimination, as well as the corresponding fairness scores, i.e., metrics that are used to quantify the degree of unfairness.

Group fairness, also known as statistical fairness, focuses on certain protected groups such as ethnic minority and the parity across different groups based on some statistical measurements. It is the primary focus of this study as well as many existing studies [7, 24, 32, 42, 49]. Classic measurements for group fairness include positive classification rate and true positive rate. A classifier satisfies group fairness if the samples in the protected groups have an equal or similar positive classification probability or true positive probability.

Given a model, we can measure its degree of unfairness according to group fairness using Statistical Parity Difference (SPD) [10][1].

Definition 2.1 (Statistical Parity Difference). Let Y be the predicted output of the neural network N; l be a (favorable) prediction and F be a protected attribute. Statistical Parity Difference is the difference in the probability of favorable outcomes between the unprivileged and privileged groups where the unprivileged/privileged groups are defined based on the value of the protected attribute.

$$|P[Y = l \mid F = 0] - P[Y = l \mid F = 1]| \qquad (1)$$

Note that the above definition only considers a single binary protected attribute, which is sometimes insufficient. The following metric, called Group Discrimination Score (GDS), extends SPD to measure fairness based on multiple protected attributes.

Definition 2.2 (Group Discrimination Score). Let N be a neural network; Y be the predicted output of the neural network; l be a (favorable) prediction, and F be a set of (one or more) protected attributes. Let θ (and θ') be an arbitrary valuation of the protected attributes F. Let X_θ be the set of inputs whose F-attribute values are θ. Let P_θ be $P(N(x) = l \mid x \in X_\theta)$. The multivariate group discrimination with respect to protected attributes F is the maximum difference between any P_θ and $P_{\theta'}$.

Example Consider the structured dataset Adult Income [38]. It has two protected attributes, i.e., gender, and race. Each attribute has a set of two values, i.e., Female or Male for gender, and White or non-White for race. As a result, there are 4 possible θ, i.e., (Male, White), (Female, White), (Male, non-White) and (Female, non-White). The probabilities of an individual who is predicted to have a high-income level (i.e., more than 50K) with respect to these four θ is 14.4%, 39.6%, 9.0% and 28.5% respectively. The GDS of the model is thus 30.6%.

Individual discrimination is another concept which is often applied in fairness analysis. It focuses on specific pairs of individuals. Intuitively, individual discrimination occurs when two individuals that differ by only certain protected attribute(s) are predicted with different labels. An individual whose label changes once its protected attribute(s) changes is referred to as an individual discriminatory instance.

Definition 2.3 (Individual Discriminatory Instance). Let F be a set of (one or more) protected attributes; and N be a neural network. x is an individual discriminatory instance if there exists an instance x' such that the following conditions are satisfied.

- $\forall q \notin F.\ x_q = x'_q$
- $N(x) \neq N(x')$

The above definition is often adopted in fairness testing, i.e., works on searching or generating individual discriminatory instances [44, 50]. In addition, there are proposals on learning models which are more likely to avoid individual discriminatory [40].

Given a model, we can measure its fairness according to individual discrimination by measuring the percentage of individual discriminatory instances in a set of instances (which can be the test set or a set generated to simulate unseen samples), formally called Causal Discrimination Score (CDS).

Definition 2.4 (Causal Discrimination Score). Let N be a neural network; F be a set of protected attributes. The causal discrimination score of N with respect to protected attributes F, is the fraction of inputs which are individual discrimination instances.

2.2 Fairness Improving Methods

Many methods have been proposed to improve the fairness of neural networks [3, 4, 11, 12, 19, 23, 28–31, 37, 46, 47]. They can be categorized into three groups according to when they are applied, i.e., pre-processing, in-processing and post-processing.

Pre-processing methods aim to reduce the discrimination and bias in the training data so as to improve the fairness of the trained model. Among the many pre-processing methods [11, 19, 28, 46], we focus on the following two representatives in this work.

- Reweighing (RW) [28] works by assigning different weights to training samples in order to reduce the effect of data biases. In particular, lower weights are assigned to favored inputs which have a higher chance of being predicted with the favorable label and higher weights are assigned to deprived inputs.
- Disparate Impact Remover (DIR) [19] is based on the disparate impact metric which compares the proportion of individuals that are predicted with the favorable label for an unprivileged group and a privileged group. It modifies the values of the non-protected attribute to remove the bias from the training dataset.

In-processing methods modify the model in different ways to mitigate the bias in the model predictions [3, 4, 12, 30, 31, 47]. We focus on the following representative in-processing methods in this work.

- Classification with fairness constraints (META) [12] develops a meta-algorithm which captures the desired metrics of group fairness (e.g., GDS), using convex fairness constraints (with strong theoretical guarantees) and then using the constraints as an additional loss function for training the neural network.
- Adversarial debiasing (AD) [47] modifies the original model by including backward feedback for predicting the protected attribute. It maximizes the predictors' ability for classification while minimizing the adversary's ability to predict the protected attribute to mitigate the bias.
- Prejudice remover regularizer (PR) [30] focuses on the indirect prejudice. It uses regularizers to compute and restrict the effect of the protected attributes.
- Exponential gradient reduction (GR) [3] reduces the fair classification problem to a sequence of cost-sensitive classification problems, whose solutions yield a randomized classifier with the lowest empirical error subject to the desired constraints.

Post-processing methods modify the prediction results instead of the inputs or the model. We consider three representative processing algorithms in this work.

- Equalized Odds (EO) [23] solves a linear program to find probabilities with which to change the output labels, so as to optimize equalized odds on protected attributes.
- Calibrated Equalized Odds (CEO) [37] optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.
- Reject Option Classification (RO) [29] assigns favorable labels to unprivileged instances and unfavorable labels to privileged instances around the decision boundary with the highest uncertainty.

Table 1: Dataset Privileged Groups Definition

Dataset	protected Attribute	Privileged Group	Favorable Class
Adult Income	gender race	gender=Male race=Caucasian	income>50K
German Credit	gender age	gender=Male age>30	good credit
Bank Marketing	age	age>30	Yes
COMPAS	gender race	gender=Female race=Caucasian	no recidivism

3 AN EMPIRICAL STUDY

In this section, we present an empirical study which aims to compare the performance of different fairness improving methods on different models, different protected attributes or attribute combinations.

3.1 Experimental Setup

Datasets Our experiments are based on 4 models trained with the following benchmark datasets: Census Income [38], German Credit [25], Bank Marketing [35] and COMPAS [6]. These datasets have been used as the evaluation subjects in multiple previous studies [15, 21, 34, 40, 49, 50].

- Adult Income: The prediction task of this dataset is to determine whether the income of an adult is above $50,000 annually. The dataset contains more than 30,000 samples. The attributes *gender*, *race* are protected attributes.
- German Credit: This is a small dataset with 600 samples. The task is to assess an individual's credit based on personal and financial records. The attributes *gender* and *age* are protected attributes.
- Bank Marketing: The dataset contains more than 45,000 samples and is used to train models for predicting whether the client would subscribe a term deposit. Its only sensitive attribute is *age*.
- COMPAS: The COMPAS Recidivism dataset contains more than 7,000 samples and is used to predict whether the recidivism risk score for an individual is high. The attributes *gender*, *race* are protected attributes.

In our experiment, we define privileged and unprivileged groups based on the default setting in [7]. The details of the privileged group definitions and favorable class are summarised at Table 1. Altogether, we have a total of 10 model-attribute combinations. Our implementation of the 9 fairness improving methods is based on the AIF360 implementation [7]. Each implementation is manually reviewed and tested through standard practice.

Model Training Our models are feed-forward neural networks, which are shown to be highly accurate and efficient in these real-world classification problems [1, 26, 48]. All these neural networks contain five hidden layers, each of which contains 64, 32, 16, 8 and 4 units. The output layer contains 2 (number of predict classes) units. For each dataset, we split the data into 70% training data and 30% test data. All experiments are conducted on a server running Ubuntu 1804 operating system with 1 Intel Core 3.10GHz CPU, 32GB memory and 2 NVIDIA GV102 GPU. To mitigate the effect of randomness, whenever relevant, we set the same random seed

Table 2: Neural Networks in Experiments

Dataset	Protected Attribute	SPD	GDS	CDS	Accuracy
Adult Income	gender	0.249	0.249	0.103	
	race	0.119	0.119	0.117	81.7%
	gender+race	-	0.306	0.179	
German Credit	gender	0.031	0.031	0.078	
	age	0.095	0.095	0.15	63.3%
	gender+age	-	0.133	0.172	
Bank	age	0.047	0.047	0.014	90.0%
COMPAS	gender	0.227	0.227	0.076	
	race	0.151	0.151	0.028	72.7%
	gender+race	-	0.301	0.083	

Table 3: Best Method for Group Fairness Improvement

Dataset	Protected Attribute	Group Fairness	Absolute Change
Adult Income	gender	GR	0.248
	race	META	0.095
	gender, race	GR	0.272
German Credit	gender	RW	0.023
	age	RW	0.101
	gender, age	RW	0.078
Bank	age	RW	0.041
COMPAS	gender	RO	0.188
	race	RO	0.14
	gender, race	RO	0.222

for each test. The trained models reach standard state-of-the-art accuracy. The trained results including the corresponding fairness scores are shown in Table 2. Note that SPD is the probability difference between the unprivileged and privileged groups which is defined on a single protected attribute and thus it is irrelevant if multiple protected attributes are considered simultaneously.

3.2 Evaluation Results

In the following, we present the results of the empirical study, which aims to answer the following research questions.

RQ1: Do the fairness improving methods always improve group fairness? To answer the question, we systematically apply all fairness improving methods on all the model-attribute combinations and measure the effectiveness of the fairness improving methods. We measure the group fairness improvement as follows. SPD is adopted if a single protected attribute is relevant and GDS is adopted if multiple protected attributes are considered at the same time. Note that GDS is the same as SPD with respect to a single protected attribute.

The results are shown in Figure 1, where there is one bar for each model-attribute combination and for each fairness improving method, i.e., a total of 9 bars for each model-attribute combination (e.g., *Adult-gender*) and 90 bars in total. A positive value means improved fairness and a negative value means worsened fairness. This bar is shown in 9 different colors for the nine different methods.

First of all, to our surprise, the fairness improving methods are not always helpful in terms of improving fairness. As shown in Figure 1, while many methods have a positive effect in many cases, *there are many instances where applying fairness improving method results in worsened fairness, sometimes quite significantly.* This is shown as the colorful bar before the zero line, which accounts for a total of 18 cases (i.e., 20%). Most of those cases are for in-processing and post-processing methods.

Furthermore, the performance of the methods varies significantly across different models and protected attributes. Table 3 shows a summary on which method achieves the most fairness improvement for each model-attribute combination and it can be observed that different winners are there for different model-attribute combinations. Further analysis shows the performance of the fairness improving methods vary across many dimensions. First, the performance of the same method varies significantly on different models.

For instance, while the post-processing method CEO works effectively for the neural network trained on Adult Income dataset, it is ineffective for the model trained on German Credit dataset. Secondly, the performance of the methods varies across different attributes in the same model. For instance, the post-processing method EO improves the group fairness with respect to *gender* attribute effectively but leads to worse group fairness with respect to *race* attribute for the neural network trained on Adult Income dataset.

Moreover, even the processing methods in the same category behave differently on the same model-attribute combination. In terms of in-processing methods, RW is much more effective than DIR. All models' group fairness can be improved by RW, whereas DIR is ineffective with respect to *Credit-gender* and *COMPAS-race*. For in-processing methods, GR is most effective in improving group fairness for all model-attribute combinations except *Credit-age.* The performance of Post-processing methods varies significantly. For example, the post-processing method RO is much more effective in improving the group fairness for the neural network trained on COMPAS dataset than CEO and EO.

There are some conjectures on why fairness improvement approaches may have different effects on different models and different model-attribute combinations. The main reason is that these methods improve fairness based on certain metrics which may be subtly different from common notions of fairness such as SPG, GDS and CDS. For instance, CEO focuses on reducing False Positive Rate difference in particular, which sometimes translates to fairness measured using SPG/GDS/CDS (as for the Adult Income dataset) and sometimes not. For the different performances on different model-attribute combinations, there may be two reasons. The first is that the discrimination against different attributes in the model may be very different (see in Table 2 and observed in [8]). The second possible reason is that the reasons of the discrimination against different attributes may be different, e.g., biased training data or biased models.

> **Answer to RQ1: Existing fairness improving methods are not always effective in improving group fairness and thus they must be applied with caution.**

RQ2: What is the cost on accuracy when applying existing fairness improving methods? The results are shown in Figure 2, where there is similarly one bar for each model-attribute combination and for each
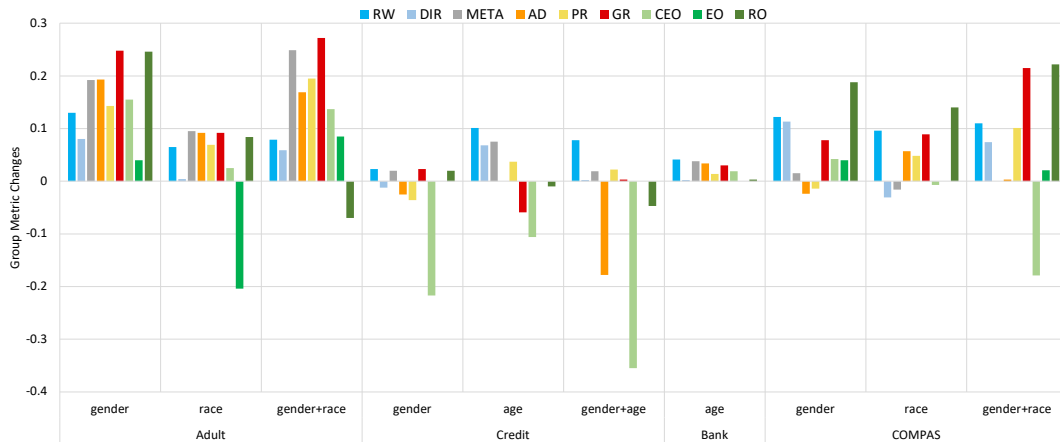
Figure 1: Group Fairness Improvement of Models with respect to Different Protected Attributes
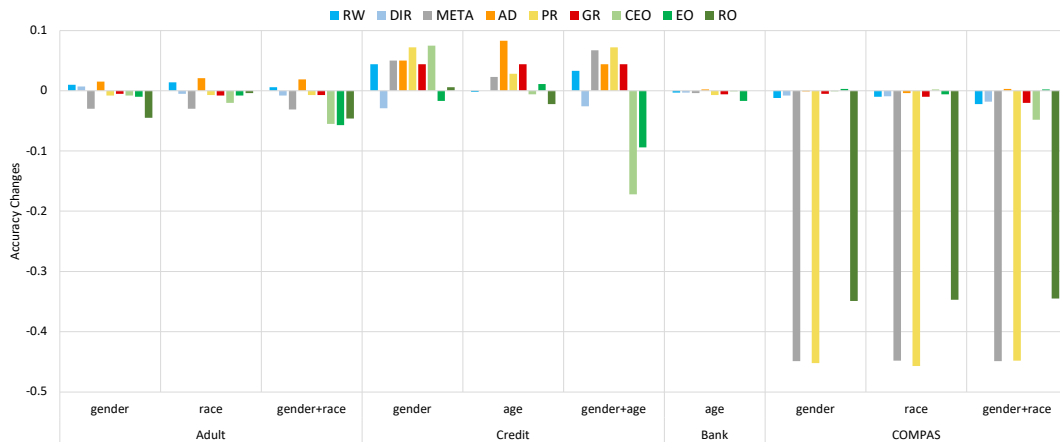

Figure 2: Accuracy Changes of Models with respect to Different Protected Attributes After Processing

fairness improving method. A positive value indicates an increased accuracy and a negative value indicates a decreased accuracy.

First of all, we observe that some of the fairness improving methods may indeed incur a significant loss of accuracy. This is most observable on META, PR, CEO, EO and RO. Especially for the neural network trained on the COMPAS dataset, the accuracy drops more than 40% after applying META, PR or RO. The average loss of accuracy is around 13% after processing by META and 12% after processing by RO. To our surprise, some of the fairness improving methods result in improved accuracy in some cases. This is most observable in some in-processing methods. Especially for the neural network trained on the German Credit dataset, the accuracy increases after applying all four in-processing methods. It should be noted however most of these in-processing methods have a less or harmful effect in terms of group fairness improvement in these cases. For example, while the accuracy increases by 4% after applying GR on *Credit-age*, the SPD fairness score worsens by 6%.

The accuracy reduction varies across not only different model-attribute combinations, but also different methods across different categories. Compared fairness improving methods from different categories, the pre-processing methods have an overall mild impact on the model accuracy. In terms of the most effective pre-processing method RW, it is effective on group fairness improvement with respect to all model-attribute combinations and scarifies little accuracy. In terms of the most effective in-processing method GR, it is effective on group fairness improvement with respect to all model-attribute combinations except *Credit-age* (although sometimes with minimal fairness improvement). Among them, 7 neural networks get lower accuracy after processing. But the accuracy drops less than 1% in average. In terms of the post-processing method RO, it is effective on group fairness improvement with respect to 7 model-attribute but 5 neural networks get lower accuracy after processing. Especially for the neural network trained on COMPAS dataset, the accuracy drops more than 30%, which is unacceptable.

> **Answer to RQ2: Existing fairness improving methods may incur a significant loss in accuracy.**
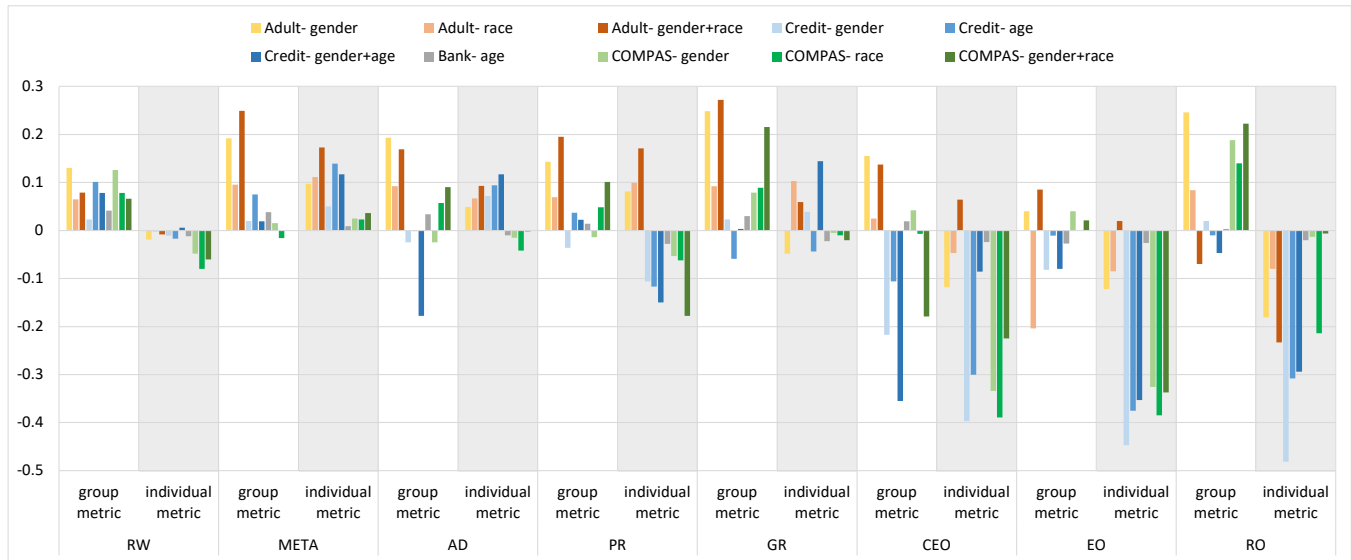
Figure 3: Comparison between group fairness improvement and individual discrimination reduction

RQ3: Do the fairness improving methods perform differently for improving group fairness and reducing individual discrimination? Almost all existing fairness improving methods focus on group fairness (whilst some fairness testing approaches focus on individual discrimination for some reason). Thus we are curious about whether the existing fairness improving methods can reduce individual discrimination as well. To answer this question, we compare the CDS change against the group fairness metric change achieved by the same method. The idea is to check whether the changes are consistent, i.e., whether an improvement in group fairness leads to a reduction in individual discrimination and vice versa. Note that, by the default setting in [7], the DIR pre-processing method removes all protected attributes, which makes individual discrimination irrelevant, and thus is not considered in this experiment.

The results are shown in Figure 3, where the CDS change is placed next to the fairness metric change for each fairness improving method. First of all, the group fairness improvement and individual discrimination reduction are inconsistent. A method improving the group fairness effectively might have none or even harmful effect on individual fairness. This is most observable on RW and RO. The pre-processing method RW is effective on group fairness improvement for all models but lead to more individual discrimination for 8 model-attribute combinations. After applying the post-processing method RO, the individual discrimination worsens for all model-attribute combinations.

Furthermore, only the in-processing methods consistently reduce individual discrimination. In terms of META method, it increases the group fairness and reduces the individual discrimination at the same time for 8 model-attribute combinations. The method AD reduces the individual discrimination with respect to all protected attributes in Adult Income dataset and German Credit dataset. Especially for the neural network trained on Adult Income dataset, all in-processing methods improve the individual fairness effectively. By contrary, all post-processing methods have harmful effect on

individual discrimination. on average, the CDS worsens by around 19% after applying CEO, worsens by 24% after applying EO and worsens by more than 18% with RO.

> **Answer to RQ3: Existing methods are less effective in reducing individual discrimination.**

4 AN ADAPTIVE APPROACH

Our empirical study shows that the performance of fairness improving methods varies significantly across different models, i.e., sometimes resulting in worsened fairness and/or reduced accuracy. We thus need a systematic way of choosing the right method. Our proposal is an adaptive approach based on causality analysis. Intuitively, causality analysis measures the "responsibility" of each neuron and input attributes towards the unfairness, and depending on whether the most responsible neurons are in the hidden layers or at the input layer, as well as whether a small number of them are significantly more responsible than the rest. Then we choose the fairness improving method accordingly. In the following, we present the details of our approach.

4.1 Causality Analysis

Causality analysis aims to identify the presence of causal relationships among events. Furthermore, it can be used to quantify the causal influence of an event on another event. To conduct causality analysis on neural networks, we first adopt the approach in [14, 42], and treat neural networks as Structured Causal Models (SCM). Formally,

Definition 4.1 (Structure Causal Model). A Structure Causal Model consists of a set of endogenous variables X and a set of exogenous variables U connected by a set of functions F that determine the values of the variables in X based on the values of the variables

in U. The neural network corresponding SCM can be represented as a 4-tuple Model $M(X, U, F, P_U)$, where P_U is the probability of distribution over U. □

For the neural network, the endogenous variables V are observed variables, e.g., attributes or neurons. The exogenous variables are the unobserved random variables, e.g., noise, and P_U is the possible distribution of the exogenous variables. Trivially, an SCM can be represented by a directed graphical model $G = (X, E)$, where E is the causal mechanism.

Based on SCM, the causal effect of a certain event can be computed as the difference between potential outcomes under different treatments. In this work, we adopt the Average Causal Effect (ACE) as the measurement of the causal effect [14, 42][2]. The formal definitions of ACE are shown below (where it is assumed that the input endogenous variables are not correlated to each other).

Definition 4.2 (Average Causal Effect). The ACE of a given endogenous variable x with value α on output y can be measured as:

$$ACE_{do(x=\alpha)}^{y} = \mathbb{E}[y \mid do(x = \alpha)] - baseline_x \qquad (2)$$

where $\mathbb{E}[y \mid do(x_i = \alpha)]$ represents the interventional expectation which is the expected value of y when the random variable x is set to α; and $baseline_x$ is the average ACE of x on y, i.e., $\mathbb{E}_x\left[\mathbb{E}_y[y \mid do(x = \alpha)]\right]$[3]. □

Following the recent work reported in [42], we apply ACE to capture the causal influence on model fairness. That is, the y in Equation 2 should be a measure of the model unfairness, i.e., SPR, GDS or CDS. For simplicity, we denote it as y_{fair}.

In order to analyze the causal effect on fairness, we analyze two possible causal effects, i.e., the relationship between input attributes to unfairness, and the relationship between the hidden neurons to unfairness. In this work, we make use of the average interventional expectation to approximate the ACE of variable x to y_{fair}. Formally, $ACE_{do(x=\alpha)}^{y_{fair}}$ represents the ACE of variable x under value α to the fairness property y_{fair}. One complication is that each input attribute or neuron has many possible values and we must consider all the possible values in computing the ACE. Our remedy is to consider the average Interventional Expectation (AIE).

Definition 4.3 (Average Interventional Expectation). Let x be the given endogenous variable, y_{fair} be the fairness property and val_set_x be a set of values of variable x. The average interventional expectation is the mean of expected values of y_{fair} when x is set to be each value α:

$$AIE_x^{y_{fair}} = \frac{\sum_{\alpha \in val\_set_x} \mathbb{E}[y_{fair} \mid do(x = \alpha)]}{\#(val\_set_x)} \qquad (3)$$

For the input features with categorical values, we intervene the feature with every possible value based on the training dataset. For the hidden neurons with continuous value, intervening it with every possible value might be consuming. We thus intervene the neurons as follows which is adopted in [14] as well. That is, we assume the "intervener" is equally likely to perturb variable x to any value α within the input range, so that $\alpha \sim U(min_x, max_x)$, where

---

Algorithm 1 $CausalityNeuron(N, D, n, fair\_metric)$ where N is the neural network, D is the dataset used to measure causal effect, n is a hidden neuron in N and $fair\_metric$ is the function measuring the fairness score based on the desired fairness metric

1: $min :=$ minimum output of neuron n
2: $max :=$ maximum output of neuron n
3: $val\_set = generate\_vals(min, max, num\_interval)$
4: **for** α in val_set **do**
5: $\quad ie \leftarrow \{\}$
6: $\quad y_{fair} = fair\_metric(N, D|do(n = \alpha))$
7: $\quad ie \leftarrow ie \cup y_{fair}$
8: **end for**
9: **return** $mean(ie)$

Algorithm 2 $CausalityAttribute(N, D, f, fair\_metric)$ where N is the neural network, D is the dataset used to measure causal effect, f is an input attribute and $fair\_metric$ is the function of measuring the fairness score based on the desired fairness metric

1: $val\_set :=$ the set of all possible values of attribute f
2: **for** α in val_set **do**
3: $\quad ie \leftarrow \{\}$
4: $\quad y_{fair} = fair\_metric(N, D|do(f = \alpha))$
5: $\quad ie \leftarrow ie \cup y_{fair}$
6: **end for**
7: **return** $mean(ie)$

---

min_x and max_x are the minimum and maximum input values of x. In practice, min_x and max_x can be obtained by observing the value of the input attribute or neuron given all the training samples and the val_set_x is generated by partitioning the range $[min_x, max_x]$ uniformly into a fixed number of intervals. Note that if a specific distribution of the interventions is given, it can be used to generate the intervention values instead.

The details of causality analysis on the hidden neurons are shown in Algorithm 1. Given a neural network N, a set of inputs D (i.e., the training set), a hidden neuron n and the function for measuring the desired fairness score $fair\_metric$, we systematically measure the AIE with neuron intervention. At line 1 and line 2, we set min to the minimum output of n and max to the maximum output of n. Then we generate a set of evenly spaced numbers within the domain of the neuron output $[min, max]$ as val_set through function $generate\_vals$ at line 3. The input parameter $num\_interval$ decides how many intervals are there. From line 4 to 8, we calculate the AIE with each perturbing value α. In each round, we first set ie as an empty set at line 5 and then calculate the fairness score y_{fair} whilst fixing the value of neuron n as α. At line 9, we return the mean of all Interventional Expectation as the AIE.

Algorithm 2 similarly conducts causality analysis on the input attributes. The only difference is that we perform the intervention on the given attribute f at line 4 with all possible values of the attribute.

4.2 Adaptive Fairness Improvement

Once we compute the causal effect of each neuron and each input attribute on fairness (i.e., responsibility for unfairness), we can then

adaptively select the fairness improving methods. For example, if the causal effects of input attributes are relatively high, the unfairness is more likely to be related to the input attributes and likely to be eliminated by pre-processing methods. Similarly, if the interior neurons in the neural network have high causal effects on the fairness property, in-processing methods might be a suitable choice for fairness improvement.

Formally, to properly compare the casual effects of neurons and input attributes, we first normalize it with respect to a baseline $\overline{y_{fair}}$, which is the fairness score based on the desired fairness metric without any intervention. The baseline $\overline{y_{fair}}$ can be SPD, GDS and CDS as discussed previously.

We define the causal effects higher than the basic fairness property as high causal effects and vice versa. In other words, only the variable with a causal effect higher than the basic fairness property has the positive causality to unfairness. That is, we only consider those neurons and attributes with a causal effect higher than $\overline{y_{fair}}$ as responsible. Next, we measure the proposition of input attributes and neurons that are considered responsible. Given the set of causal effects of all attribute AIE_f and the set of causal effects of all neurons AIE_n, we formally denote the proportion of high causality attributes as P_f and the proportion of high causality neurons as P_n and define them as follows.

$$P_f = P(AIE > \overline{y_{fair}} \mid AIE \in AIE_f) \qquad (4)$$

$$P_n = P(AIE > \overline{y_{fair}} \mid AIE \in AIE_n) \qquad (5)$$

Furthermore, we measure the distribution of the "responsibility" among the input attributes and neurons, since it intuitively has an impact on which fairness improving method should be chosen. For instance, if all input attributes have similar responsibility for unfairness, it is likely hard to pre-process the inputs so as to eliminate the discrimination. Similarly, if all neurons are equally responsible for unfairness, it is complex to improve the fairness by focusing on a few neurons as in [42]. Formally, we use the Coefficient of Variation (CV) to capture the distribution of the causal effects. CV is used to measure the dispersion of data points around the mean. It represents the ratio of the standard deviation to the mean which indicates the degree of variation. In this setting, the larger the CV, the more uneven the distribution of causal effects. We denote the CV of attributes as CV_f and the CV of neurons as CV_n.

The details of how to select fairness improving methods are shown in algorithm 3. If both the proportion of responsible attributes and responsible neurons is less than a proportion threshold P_thres, few input attributes and neurons are to be blamed for the unfairness. As a result, it is unlikely pre-processing (which focuses on input attributes) or in-processing (which focuses on the hidden neurons) is effective, and thus we choose to apply the post-processing methods. In practice, we set the threshold P_thres to be 10%. Otherwise, there are sufficient number of input attributes or neurons that are responsible for unfairness, we then select to apply a pre-processing method if $CV_f > CV_n$, i.e., the distribution of causal effects is more uneven in the input attributes which means that some of the input attributes are more responsible. Otherwise, an in-processing method is chosen. For pre-processing methods, RW is preferred over DIR, as RW is also feasible to individual fairness metrics. For in-processing methods and post-processing

---

Algorithm 3 $AdaptiveImprove(P_n, P_f, CV_n, CV_f)$

1: **if** $P_f \leq P\_thres$ and $P_n \leq P\_thres$ **then**
2:     **return** post-processing methods
3: **else**
4:     **if** $CV_f > CV_n$ **then**
5:         **return** pre-processing methods
6:     **else**
7:         **return** in-processing methods
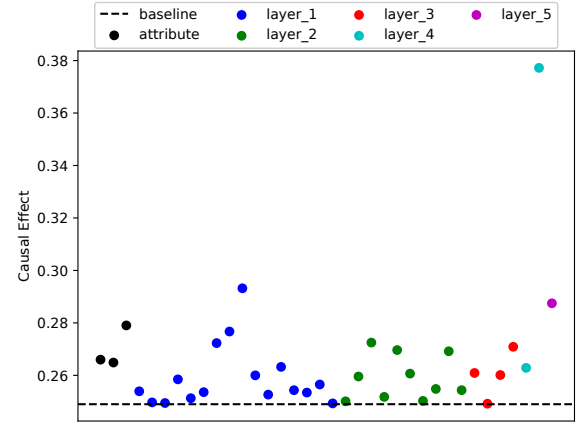8:   **end if**
9: **end if**

Figure 4: Causality analysis result of *Adult-gender*

methods, we choose the method with the best improvement and least accuracy cost.

Example For the neural network trained on Adult Income dataset, assume that the protected attribute is the "gender" attribute. According to the above discussion, we use the group fairness metric SPD to calculate the causal effects of attributes and neurons. The causality analysis result is shown in Figure 4, where each dot represents the AIE of either an input attribute or a hidden neuron. We mark the causal effects of input attributes with black dots and mark the causal effects of hidden neurons in different layers with different colors. The dotted line marks the baseline $\overline{y_{fair}}$ which is 0.249. There are 3 (i.e., 25%) attributes with causal effects higher than the baseline and 33 (i.e., 26.6%) neurons with causal effects higher than the baseline. As the proportion of responsible input attributes and neurons satisfy the threshold, we then calculate the CV values of those responsible attributes and neurons. The CV_f of these 3 attributes is 0.041 and the CV_n of these 33 neurons is 0.152. Since $CV_n > CV_f$, we choose to apply in-processing methods so as to improve the model's group fairness.

5 IMPLEMENTATION AND EVALUATION

In this section, we evaluate the performance of our adaptive approach systematically to answer multiple research questions. Note that the same datasets, models, and the configuration from Section 3

are used in this section.

RQ1: How are the "responsibility" distributed among the neurons and input attributes To answer this question, we show the probability of high causal effects and CV of these causal effects for both the hidden neurons and input attributes in Table 4 and Table 5. The first column is the training dataset and the second column shows the corresponding protected attribute(s) in each dataset. Then we show the probability of attributes with high causal effects P_f, the probability of neurons with high causal effects P_n, CV on highly causal attributes CV_f and CV on highly causal neurons CV_n. It can be observed that the distribution of responsibility varies significantly across different model-attribute combinations, which potentially explains why only some fairness improving methods are effective sometimes.

Table 4 shows the distribution of high causal effects based on group fairness metrics, e.g., SPD for single protected attributes and GDS for multivariate protected attributes. Based on algorithm 3, the selected processing categories are shown in the last column. For all attribute(s) in Adult Income dataset, the probabilities of high causal effects are higher than 10% and CV_n scores are higher than CV_f scores. So we decide to apply pre-processing methods to this model to improve the group fairness for all attributes. For the neural network trained on German Credit dataset with respect to all attributes, we conclude to apply pre-processing methods. For example, with respect to *age* attribute, both the proportion and the CV of high causal neurons are lower than the two of high causal attributes. Similarly, based on the distribution of high causal effects, we conclude to apply pre-processing to the neural network trained on Bank dataset and the neural network trained on COMPAS dataset with respect to *gender* and *race* attributes. With respect to *gender+race* attribute in COMPAS dataset, as the CV of neurons is higher, we conclude to apply in-processing methods.

Table 5 show the distribution of high causal effects based on individual fairness metrics, e.g., CDS. The selected processing categories are shown in the last column. Similarly, Algorithm 3 decides to apply in-processing methods for all model-attribute combinations, expect *Credit-gender* and *Bank-age*. We can observe that the proportion of high causal effects of attributes might be 0% in some cases, e.g., *COMPAS-gender* and *COMPAS-race*, which means no attribute is responsible for individual discrimination.

Note that, post-processing methods are selected only if both the proportions of responsible neurons/attributes are low, as it often has a significant negative impact on model performance (so that it is impossible to improve fairness through pre-processing or in-processing). In our experiments, however, all the neural networks have sufficiently many responsible neurons/attributes, so no post-processing method is adopted.

RQ2: Are we always able to identify the best performing fairness improvement method? To answer this question, we compare our adaptive approach against the best performing pre-processing, in-processing and post-processing method in four ways.

- One is the group fairness improvement, which is shown in Figure 5(a).

Table 4: Distribution of high causal effects with Group Fairness

Dataset	Protected Attribute	P_f	P_n	CV_f	CV_n	Processing
Adult Income	gender	25.0%	26.6%	0.041	0.152	in-processing
	race	16.6%	28.2%	0.104	0.215	in-processing
	gender+race	27.3%	26.7%	0.095	0.163	in-processing
German Credit	gender	73.7%	46.0%	0.339	0.323	pre-processing
	age	21.1%	9.6%	0.160	0.096	pre-processing
	gender+age	77.8%	53.2%	0.269	0.235	pre-processing
Bank	age	33.3%	37.9%	0.183	0.142	pre-processing
COMPAS	gender	63.6%	43.5%	0.052	0.045	pre-processing
	race	36.4%	19.4%	0.056	0.034	pre-processing
	gender+race	60.0%	86.3%	0.0018	0.002	in-processing

Table 5: Distribution of high causal effects with Individual Discrimination

Dataset	Protected Attribute	P_f	P_n	CV_f	CV_n	Processing
Adult Income	gender	75.0%	58.8%	0.033	0.058	in-processing
	race	75.0%	38.7%	0.128	0.141	in-processing
	gender+race	63.3%	46.8%	0.091	0.105	in-processing
German Credit	gender	94.7%	70.2%	0.114	0.096	pre-processing
	age	63.2%	29.0%	0.041	0.053	in-processing
	gender+age	83.3%	10.3%	0..061	0.066	in-processing
Bank	age	40.0%	50.8%	0.076	0.047	pre-processing
COMPAS	gender	0%	15.3%	-	0.026	in-processing
	race	0%	21.0%	-	0.133	in-processing
	gender+race	30%	39.5%	0.075	0.1	in-processing

- One is the group fairness improvement minus the accuracy loss, which is shown in Figure 5(b).
- One is the individual discrimination reduction, which is shown in Figure 6(a).
- One is the individual discrimination reduction minus the accuracy loss, which is shown in Figure 6(b).

As shown in Figure 5(a), if we focus on group fairness improvement only, our approach achieves the best performance for 7 out of 10 cases, e.g., *e.g., all attributes in Adult Income dataset, all attributes in German Credit dataset the attribute in Bank dataset*. Although for the neural network trained on Compas dataset, our adaptive approach does not have the best fairness improvement. If we consider at the same time the accuracy loss, as shown in Figure 5(b), our approach performs the best in all of the cases. Note that while the post-processing method RO often improves the group fairness significantly, the accuracy often drops significantly (e.g., more than 30% after processing with respect to all protected attributes for the COMPAS dataset, which is clearly unacceptable). In fact, according to our experiments, post-processing should rarely be the choice if we would be maintain high-accuracy.The results shown in Figure 5(b) clearly suggests that our approach is able to improve fairness effectively whilst maintaining a high accuracy.

In Figure 6(a), we show the comparison between our approach and the existing approaches in terms of reducing individual discrimination. We can observe that only the in-processing methods can reduce the individual discrimination effectively. In fact, our Adaptive Processing Algorithm 3 almost always selects to apply in-processing methods, except for *Credit-gender* and *Bank-age*. After applying the in-processing method RW, the CDS remains almost
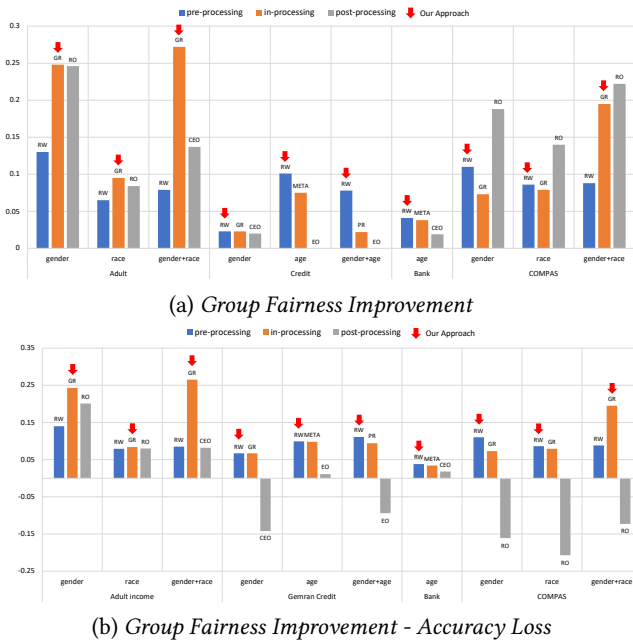
(a) *Group Fairness Improvement*

(b) *Group Fairness Improvement - Accuracy Loss*

Figure 5: Our Approach vs SOTA on Group Fairness

Table 6: Time overhead for causality analysis

Dataset	Protected Attribute	Time(s)
Adult Income	gender	495.26
	race	504.72
	gender, race	553.42
German Credit	gender	107.79
	age	116.56
	gender, age	221.72
Bank	age	550.52
COMPAS	gender	106.37
	race	152.42
	gender, race	162.19

(a) *Individual Discrimination Reduction*

(b) *Individual Discrimination Reduction - Accuracy Loss*
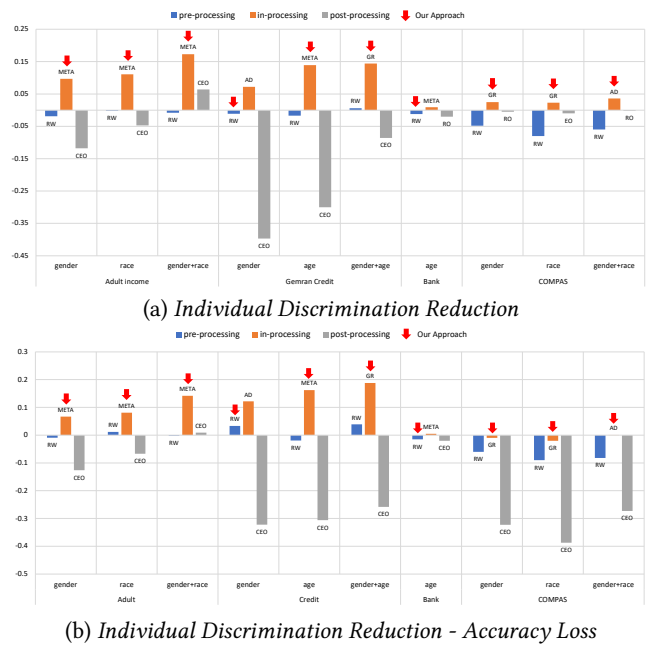
Figure 6: Our Approach vs SOTA on Individual Discrimination

the same with respect to *Credit-gender* but worsens by around 2% with respect to *Bank-age*. Taking accuracy loss into account at the same time, we show the individual discrimination reduction minus the accuracy lost in Figure 6(b). Our approach performs best in 8 out of 10 cases, except for the two cases where RW is selected for *Credit-gender* and *Bank-age*. One potential reason why this is the case is that existing pre-processing methods are not designed for reducing individual discrimination and as a result, even if a small number of input attributes are indeed responsible for the unfairness, existing pre-processing methods such as RW are not able to remove biases in the training set effectively. This calls for research into alternative pre-processing methods for reducing individual discrimination.

It is worth noting that with our approach, we always (10 out 10) achieve improved group fairness and almost always (9 out 10) achieve reduced individual discrimination, whist achieving a low accuracy loss.

RQ3: What are the time overhead for causality analysis? The time spent on causality analysis is summarised in Table 6. Note that the time is the additional time a user has to spend on applying our method before applying the selected fairness improving method. The time required for causality analysis is always less than 10 minutes.

6 THREATS TO VALIDITY

Limited model structures We currently support feed-forward neural networks (for tabular data) and convolutional neural networks (for images). It is possible to extend our method to support deep learning architectures such as RNN (for text data) by extending

causality analysis to handle feedback loops. We focus on feed-forward NN as existing studies on fairness largely focus on tabular data [15, 21, 34, 40, 49, 50].

Limited fairness metrics We only use SPD and GDS metrics for group fairness and CDS metric for individual fairness. We focus on GPD and GDS as they are the primary focuses of existing works [5, 7, 24, 32, 42, 49]. Given that GPD and GDS are similar with other metrics which consider positive classification rate like Disparate Impact, our method could work for other notions of fairness as well.

Causal effect measurement ACE is commonly used to evaluate causality [14, 42]. According to [14], alternative measurements like

integrated gradients and gradients of causal effect [36] might suffer from sensitivity and induce causal effects by other input features.

Distributional shift in the data Our approach might be affected by distributional shifts in the data. We evaluate the stability of our approach against slight distributional shifts on Adult Income dataset. Firstly, following [20], we randomly split train/test set 10 times and then evaluate whther the method selected by our approach is the best one for each of the 10 test sets. Secondly, following [44], we evaluate our approach using data generated by perturbation. In both conditions, the results confirm that is the case. It shows perhaps that our approach is robust to such levels of distributional shift.

7 RELATED WORK

This work is related to research on fairness improving methods, fairness testing and fairness verification methods as well as broadly various studies on fairness. Besides those mentioned in the previous sections, we summarize other related works below.

Fairness Testing and Verification Some existing works attempted to test model discrimination with fairness score measurements. In [43], Tramer *et al.* propose an unwarranted associations framework to detect unfair, discriminatory or offensive user treatment in data-driven applications. It identifies discrimination according to multiple metrics including the CV score, related ratio and associations between outputs and protected attributes. In [33], Kleinberg *et al.* also test multiple discrimination scores and compare different fairness metrics. In [21], Galhotra *et al.* propose a tool called THEMIS to measure software discrimination. It tests discrimination with two fairness definitions, i.e., group discrimination score and causal discrimination score. It measures these two scores for different software instances with respect to race and gender separately. Their approach generates additional testing samples by selecting random values from the domain for all attributes. In [2], Adebayo *et al.* try to determine the relative significance of a model's inputs in determining the outcomes and use it to assess the discriminatory extent of the model. In [22], Ghosh *et al.* verify different fairness measures of the learning process with respect to underlying data distribution.

Empirical Studies of Fairness Chakraborty *et al.* empirically research on the effectiveness and efficiency of existing fairness improvement methods based on group fairness metrics [13]. Friedler *et al.* work on an empirical study to compare the effects of different fairness improvement methods [20]. In [8], Biswas *et al.* focus on an empirical evaluation of fairness and mitigation on 8 different real-world machine learning models. They apply 7 mitigation techniques to these models and analyzed the fairness, mitigation results, and impacts on performance. They also present different trade-off choices of fairness mitigation decisions. Zhang *et al.* discuss how key aspects of machine learning systems, such as attribute set and training data, affect fairness in [49]. Kearns *et al.* test the effectiveness and measure the trade-offs between rich subgroup fairness and accuracy in [32]. In [16], Dodge *et al.* propose four types of programmatically generated explanations to understand fairness in machine learning systems.

8 CONCLUSION

In this paper, we empirically evaluate 9 fairness improving methods on 4 real world dataset and 90 model-attribute combinations with 3 different fairness metric. Our evaluation shows that existing fairness improving methods are not always effective in improving group fairness and are often not effective in reducing individual discrimination. Meanwhile, we test the trade-off between fairness improvement and accuracy cost. Motivated by the empirical study, we propose a light weight approach to choose the the optimal fairness improving method adaptively based on causality analysis. That is, we identify on the distribution of "responsible" attribute and neurons and choose the methods accordingly. Our evaluation shows that our approach is effective in choosing the optimal improvement method.

REFERENCES

[1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon* 4, 11 (2018), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

[2] Julius Adebayo and Lalana Kagal. 2016. Iterative orthogonal feature projection for diagnosing bias in black-box models. *arXiv preprint arXiv:1611.04967* (2016).

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[4] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[5] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 871–875. https://doi.org/10.1145/3236024.3264590

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[8] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*. 642–653. https://doi.org/10.1145/3368089.3409704

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[10] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[11] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[12] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency.* 319–328. https://doi.org/10.1145/3287560.3287586

[13] Joymallya Chakraborty, Tianpei Xia, Fahmid M Fahid, and Tim Menzies. 2019. Software engineering for fairness: A case study with hyperparameter optimization. *arXiv preprint arXiv:1905.05786* (2019).

[14] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. 2019. Neural network attributions: A causal perspective. In *International Conference on Machine Learning.* PMLR, 981–990.

[15] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 67–73. https://doi.org/10.1145/3278721.3278729

[16] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces.* 275–285. https://doi.org/10.1145/3301275.3302310

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference.* 214–226. https://doi.org/10.1145/2090236.2090255

[18] Jeroen Eggermont, Joost N Kok, and Walter A Kosters. 2004. Genetic programming for data classification: Partitioning the search space. In *Proceedings of the 2004 ACM symposium on Applied computing.* 1001–1005. https://doi.org/10.1145/967900.968104

[19] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* 259–268. https://doi.org/10.1145/2783258.2783311

[20] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency.* 329–338. https://doi.org/10.1145/3287560.3287589

[21] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering.* 498–510. https://doi.org/10.1145/3106237.3106277

[22] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. 2020. Justicia: A Stochastic SAT Approach to Formally Verify Fairness. *arXiv preprint arXiv:2009.06516* (2020).

[23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[24] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 392–402. https://doi.org/10.1145/3351095.3372831

[25] Hans Hofmann. 1994. German credit dataset. (1994). https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

[26] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence* 22, 1 (2000), 4–37. https://doi.org/10.1109/34.824819

[27] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems* 29 (2016).

[28] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[29] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining.* IEEE, 924–929. https://doi.org/10.1109/ICDM.2012.45

[30] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases.* Springer, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

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

[32] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency.* 100–109. https://doi.org/10.1145/3287560.3287592

[33] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016). https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[34] Pingchuan Ma, Shuai Wang, and Jin Liu. 2020. Metamorphic testing and certified mitigation of fairness violations in nlp models. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI).* 458–465. https://doi.org/10.24963/ijcai.2020/64

[35] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. https://archive.ics.uci.edu/ml/datasets/bank+marketing. *Decision Support Systems* 62 (2014), 22–31. https://doi.org/10.1016/j.dss.2014.03.001 https://archive.ics.uci.edu/ml/datasets/bank+marketing.

[36] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms.* The MIT Press.

[37] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[38] Barry Becker Ronny Kohavi. 1996. Data Mining and Visualization. (1996). https://archive.ics.uci.edu/ml/datasets/adult.

[39] Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, and Rocky Bhatia. 2020. A Machine Learning approach for automation of Resume Recommendation system. *Procedia Computer Science* 167 (2020), 2318–2327. https://doi.org/10.1016/j.procs.2020.03.284

[40] Anian Ruoss, Mislav Balunović, Marc Fischer, and Martin Vechev. 2020. Learning certified individually fair representations. *arXiv preprint arXiv:2002.10312* (2020).

[41] Vitomir Štruc and Nikola Pavešić. 2009. Gabor-based kernel partial-least-squares discrimination features for face recognition. *Informatica* 20, 1 (2009), 115–138. https://doi.org/10.15388/Informatica.2009.240

[42] Bing Sun, Jun Sun, Long H Pham, and Jie Shi. 2022. Causality-based neural network repair. In *Proceedings of the 44th International Conference on Software Engineering.* 338–349. https://doi.org/10.1145/3510003.3510080

[43] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P).* IEEE, 401–416.

[44] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. 2018. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering.* 98–108. https://doi.org/10.1145/3238147.3238165

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

[46] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning.* PMLR, 325–333.

[47] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 335–340. https://doi.org/10.1145/3278721.3278779

[48] Guoqiang Peter Zhang. 2000. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 30, 4 (2000), 451–462. https://doi.org/10.1109/5326.897072

[49] Jie M Zhang and Mark Harman. 2021. "Ignorance and Prejudice" in Software Fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE).* IEEE, 1436–1447. https://doi.org/10.1109/ICSE43902.2021.00129

[50] Peixin Zhang, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. 2020. White-box Fairness Testing through Adversarial Sampling. *Proceedings of the 42rd International Conference on Software Engineering (ICSE 2020), Seoul, South Korea* (2020). https://doi.org/10.48550/arXiv.2107.08176