Adversarial Learning of Group and Individual Fair Representations

Hao Liu*[0000-0003-2209-117X] and Raymond Chi-Wing Wong[0000-0001-7045-6503]

The Hong Kong University of Science and Technology
{hliubs,raywong}@cse.ust.hk

Abstract. Fairness is increasingly becoming an important issue in machine learning. Representation learning is a popular approach recently that aims at mitigating discrimination by generating representation on the historical data so that further predictive analysis conducted on the representation is fair. Inspired by this approach, we propose a novel structure, called GIFair, for generating a representation that can simultaneously reconcile utility with both group and individual fairness, compared with most relevant studies that only focus on group fairness. Due to the conflict of the two fairness targets, we need to trade group fairness off against individual fairness in addition to considering the utility of classifiers. To achieve an optimized trade-off performance, we include a focal loss function so that all the targets can receive more balanced attention. Experiments conducted on three real datasets show that GIFair can achieve a better utility-fairness trade-off compared with existing models.

Keywords: Fairness · Adversarial Learning · Learning Representation.

1 Introduction

Fairness is increasingly becoming an important issue in machine learning. Many studies have shown that using unfair historical datasets that are biased against some groups of people to train accurate machine learning models for decision-making can lead to discrimination of those groups. We refer to groups that are often discriminated against as *protected groups* (e.g., women and African-Americans), and the corresponding attributes that define them as *protected attributes* (e.g., gender and race). For instance, when evaluating loan applications, a bank officer may use applicant information such as age, gender, and credit history to determine creditworthiness, leading to a lower likelihood of approval for applications from women [1]. Motivated by this, we want to propose a fair classification model to help alleviate discrimination in decision-making systems.

To assess the fairness of various classification models, many fairness notions have been proposed and most of them can be divided into *group fairness* [2, 3] and *individual fairness* [4, 5]. Group fairness requires treating different groups

* Corresponding author

defined by protected attributes equally. Individual fairness requires *similar* individuals should be treated *similarly* by classifiers. Based on these fairness notions, many approaches [6–9] have been proposed to solve the fair classification problem. Among them, representation learning [8, 9] is a common approach, which transforms the original datasets into new representations that obfuscate the information about the protected attributes in the representations. Then, different groups have similar representations and will be treated similarly by any classifier, which satisfies group fairness. However, most existing studies only focus on group fairness, which may harm individual fairness and create discrimination. For example, in hiring decision, some unqualified people in the protected group (e.g., females) are interviewed deliberately [7], which is, in fact, biased against the individuals in the unprotected group. Individual fairness can alleviate such discrimination by ensuring that individuals who are similar in terms of attributes/background (e.g., similar academic experience) are treated similarly.

Only a handful of studies on fair classification [7,10] consider both individual and group fairness in their designs. In LFR [7], a loss function is defined that combines accuracy, group fairness and individual fairness. However, the three terms are trained at the same time, but not well reconciled *at the same time*. Besides, the loss function in LFR enforces fairness *indirectly*, so the fairness performance of learned representation is not guaranteed. DualFair [10] explores an alternative formation of individual fairness called *counterfactual fairness* [5] which grant similar treatment for counterfactual samples, where a counterfactual sample of an individual x is defined to be a "synthetic" individual who is similar to x *except for* the protected attribute. However, counterfactual fairness cannot guarantee general and stronger individual-level fairness for *any* two similar individuals.

We mainly focus on reconciling accuracy and two types of fairness (i.e., group fairness and individual fairness). Due to the conflict between group and individual fairness [11], we aim to achieve a better *trade-off* between them. To solve this problem, we propose an approach called **GIFair** (for group and individual fair representations), which transforms the original dataset into a *fair representation*. To reconcile group and individual fairness in the learned representation, we use two adversaries, one for group fairness and the other for individual fairness, instead of using only one adversary in the related studies. For group (fairness) adversary, we apply an effective formation of target function, which better guarantees group fairness. For individual (fairness) adversary, we form its target function with a metric called *yNN* based on k-nearest neighbors, which addresses the explicit individual fairness of treating any similar individuals equally. We propose a well-designed training algorithm to reconcile all concepts in our structure. Compared to the existing adversarial learning studies that only consider accuracy and group fairness, we handle a more complicated problem with a better performance, e.g., we achieve a 3% improvement in accuracy and 40% improvement in group fairness on dataset COMPAS compared with baselines.

To further optimize GIFair, we propose a focal loss function so that the three targets receive more balanced attention. GIFair with focal loss function obtains even better trade-off performance (e.g., 30% improvement of group fairness under the same level of individual fairness) compared with the original GIFair.

We conduct extensive experiments on three real datasets to study the trade-off among accuracy, group fairness and individual fairness. The results show that compared with many baseline algorithms, GIFair can achieve better performance, e.g, GIFair can achieve up to 2% improvement in accuracy under the same individual fairness performance on dataset Adult.

The contributions of our work are as follows. (1) We design a novel structure of adversarial representation learning with two adversaries for group fairness and individual fairness, respectively. (2) We design a training algorithm that can well reconcile the two adversaries in our structure. Ablation analysis is conducted to show its superiority. (3) We propose a focal loss function to ensure balanced attention of two types of fairness and accuracy. (4) The experiments conducted on 3 real datasets show that GIFair can reconcile good fairness with high accuracy.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the preliminaries. Section 4 describes our solution to the fair classification problem. Then, Section 5 reports experimental results and our analysis. Finally, Section 6 concludes this paper.

2 Related Work

Most machine learning studies about fairness can be classified into *pre-processing*, *in-processing* and *post-processing*. *Pre-processing* approaches directly modify datasets to remove discrimination [6]. *In-processing* approaches modify the classifier to improve its fairness performance [7, 12]. *Post-processing* approaches directly change the predicted outcomes of the learned predictors [2].

Learning Fair Representations. Recently, *fair representation learning* [7] attracts great attention in fair machine learning, which is to learn a debiased representation so that the downstream tasks could satisfy fairness requirements. In this branch, iFair [12] considers a probabilistic mapping to the representation space to address both accuracy and individual fairness (which uses a similar fairness notion as in this paper) but fails to address group fairness as we do. DualFair [10] applies a contrastive self-supervised learning approach to obtain the representation satisfying both group fairness and counterfactual fairness. However, although LFR [7] and DualFair [10] set both group and individual fairness as targets, as mentioned in Section 1, they are not effective enough to address individual fairness. LFR [7] uses an *indirect* individual fairness formation that minimizes the deviation between each data point and its representation, and thus the individual fairness of the representation relies on the individual fairness of the original dataset, which is not always ensured. DualFair [10] focuses on counterfactual fairness but does not ensure individually fair results for *any* two similar samples. In comparison, we form our individual fairness notion based on an explicit target of treating any similar individuals equally.

Among those approaches, adversarial representation learning has been broadly explored. ALFR [8] provides a framework of learning representations that minimize the performance of the adversary which predicts the protected attribute of the representation. LAFTR [13] follows this framework to explore adversarial learning as a method of obtaining a representation to mitigate unfair prediction outcomes. IPM [14] proposes the integral probability metric adopted in an adversary such that a good theoretical guarantee on group fairness is obtained. However, all these existing methods focuses on group fairness only, while our method GIFair (following the idea of adversarial representation learning) reconciles both group and individual fairness by a novel structure of two adversaries.

3 Preliminaries

In the fair classification problem, we are given a dataset D containing N data points. The i-th data point in D, denoted by x_i where $i \in [1, N]$, has a list X of d features, i.e., $x_i \in \mathbb{R}^d$. Each x_i is also associated with an outcome attribute Y for classification and a protected attribute A representing the group membership (e.g., gender). Following [7,8,13], we assume binary outcome attribute and binary protected attribute (i.e., $Y \in \{0, 1\}$ and $A \in \{0, 1\}$). We assume that values 1 and 0 represent the protected group (e.g., females) and the unprotected group (e.g., males), respectively. We thus denote D_1 and D_0 to be the subsets of D containing all data points in the protected and unprotected group, respectively.

The basic goal of the fair classification problem is to obtain a classifier η that can predict an outcome $\eta(x_i) \in \{0, 1\}$ of data point x_i for $i \in [1, N]$ in the dataset D such that some fairness criteria are satisfied.

To achieve fairness, we follow common approaches to optimize some fairness metrics. For group fairness, we use two popular metrics, the *demographic parity gap* [3] and *equalized odd distance* [2]. Given a classifier η and dataset D, the *demographic parity gap* of η for D, denoted by $\Delta DP_D(\eta)$, is defined to be the absolute difference between the positive rate of D_0 and the positive rate of D_1 Namely,

$$\Delta DP_D(\eta) = |\frac{1}{|D_1|} \sum_{x_i \in D_1} \eta(x_i) - \frac{1}{|D_0|} \sum_{x_j \in D_0} \eta(x_j)| \tag{1}$$

The *equalized odd distance* of η for D, denoted by $\Delta EO_D(\eta)$, is defined to be the sum of the absolute difference between the true positive rate (TPR) of D_0 and the TPR of D_1, and the absolute difference between the false positive rate FPR of D_0 and the FPR of D_1. In this paper, we use $\Delta DP_D(\eta)$ as our major group fairness metric, but we also test $\Delta EO_D(\eta)$ as an alternative metric. For both $\Delta DP_D(\eta)$ and $\Delta EO_D(\eta)$, smaller values indicate better group fairness.

Individual fairness is another perspective of fairness, which requires that two similar individuals (i.e., data points) should be treated similarly in terms of the predicted outcome [4]. Consider a data point x_i. Let $\mathcal{N}_D^k(x_i)$ denote the set of k nearest neighbors of x_i in D, where k is a positive integer. Note that $\mathcal{N}_D^k(x_i)$ is computed based on the features X only (but not the protect attribute A). This is because the similarity of two individuals should be independent to A. To quantify the individual fairness, we adapt a commonly applied metric called *yNN* [7], which measures the consistency of the prediction results among similar data points. Specifically, given a classifier η, a positive integer k and dataset D, the yNN of η for D and k, denoted by $\Delta yNN_{D,k}(\eta)$, is defined to be

$$\Delta yNN_{D,k}(\eta) = 1 - \frac{\sum\limits_{x_i \in D} \sum\limits_{x_j \in \mathcal{N}_D^k(x_i)} |\eta(x_i) - \eta(x_j)|}{k \cdot N} \tag{2}$$

which captures the average difference between the predicted outcome of a data point x_i and that of a nearest neighbor x_j of x_i. This difference is 0 if x_i and x_j have the same predicted outcome and 1 otherwise. According to Equation 2, larger $\Delta yNN_{D,k}(\eta)$ indicates better individual fairness.

Moreover, we introduce the basic concept of generative adversarial network (GAN) [15]. It has two components, namely a *generator* G and a *discriminator* C. G aims at deceiving C by constructing synthetic data $G(z)$ that could match the real data distribution P_{data}. C aims at distinguishing whether the data comes from P_{data} or $G(z)$. Both components improve their ability through learning. That is, G is trained to generate $G(z)$ that cannot be distinguished from the real data, while C is trained to identify the outcome of $G(z)$ more accurately.

4 Methodology

4.1 Problem Statement

In this work, we follow adversarial representation learning to tackle the fair classification problem, which is to learn a representation Z by re-constructing the features X in the original dataset D. The learning goal is that any classifier trained on the representation Z is accurate to predict the outcome attribute Y and is also fair in terms of both group fairness and individual fairness.

Due to the conflict of group and individual fairness [11], the two fairness goals could not be satisfied simultaneously in most cases (an extended analysis on their incompatibility is given in our supplementary material [16]). We thus set our optimization goal of classifier η such that a balanced trade-off can be obtained among accuracy, group fairness and individual fairness.

4.2 Model

First proposed by [8], plenty of existing studies follow a general framework of adversarial representation learning for fair classification. This framework uses an *encoder* as the *generator* to generate the representation Z from X which aims to obfuscate the group membership and thus ensure group fairness. To achieve that, an *adversary* as the *discriminator* is set up to identify the group of the generated representation Z. By adversarial learning [15], while the adversary improves its ability of group identification, the encoder is also well trained to generate group-obfuscated representation Z. Finally, a (group) fair representation is obtained. However, this framework so far only addresses group fairness. It remains unsolved how to accommodate individual fairness into this framework.

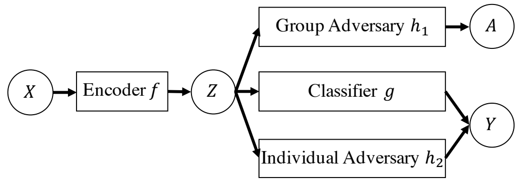

Fig. 1: Structure of GIFair

With this motivation, we propose our model called **GIFair** (**G**roup **I**ndividual **Fair**). As illustrated in Figure 1, GIFair consists of an encoder f, a classifier g and *two* adversaries, namely *group (fairness) adversary* h_1 and *individual (fairness) adversary* h_2. GIFair seeks to learn a representation Z by re-constructing the original features X of each data point in D using the encoder f. Classifier g, which predicts the outcome Y from representation Z, seeks to preserve the prediction accuracy. In addition, GIFair aims at achieving group fairness by the group adversary h_1 and individual fairness by the individual adversary h_2. Next, we introduce the details of all components and how they interact with each other.

Encoder. An encoder f: $\mathbb{R}^d \to \mathbb{R}^{d'}$ maps a data point x_i into a d'-dimensional vector, denoted by $z_i = f(x_i)$. The representation Z of the original dataset is formed by encoding all data points in D, namely, $Z = f(X) = \{f(x_i)|x_i \in D\}$.

Classifier. We use a classifier g: $\mathbb{R}^{d'} \to \{0,1\}$ to predict the outcome $g(z_i)$ of each z_i in Z and form the outcome set $g(Z) = g(f(X))$. To preserve utility, we minimize a suitable classification loss function (i.e., cross-entropy) between $g(f(X))$ and Y, denoted by $L_c(g(f(X)),Y)$ (written as L_c for simplicity).

Group Adversary. To achieve group fairness of Z, the group adversary h_1: $\mathbb{R}^{d'} \to \{0,1\}$ is included. Given a representation $z_i = f(x_i) \in Z$, h_1 generates a value $h_1(z_i) \in \{0,1\}$, which is the predicted group of z_i. Thus, we denote the set of predicted groups of Z to be $h_1(Z) = h_1(f(X))$. The objective of h_1 is to *differentiate* representations in different groups. Note that this objective *differs* from making any $h_1(z_i)$ exactly equal to the protected attribute of x_i. Instead, h_1 is only interested in giving different group labels to two representations in different groups. It is thus interesting to observe that if any $h_1(z_i)$ is wrongly predicted, h_1 also has strong differentiation performance. Therefore, following [7], we form the group (fairness) loss function on $h_1(f(X))$ and A, denoted by $L_g(h_1(f(X)),A)$ (written as L_g for simplicity), as follows.

$$L_g = L_g(h_1(f(X)),A) = |\frac{\sum\limits_{x_i \in D_0} h_1(f(x_i))}{|D_0|} - \frac{\sum\limits_{x_j \in D_1} h_1(f(x_j))}{|D_1|}| \tag{3}$$

Here, higher L_g indicates either predicting more items in D_0 as 1 and more items in D_1 as 0 (mostly wrong), or predicting more items in D_0 as 0 and more items in D_1 as 1 (mostly correct), both leading to better differentiation of representations from different groups. Thus, h_1 is trained to *maximize* L_g.

Individual Adversary. Individual fairness requires that individuals who are similar on their features X should be *indistinguishable* in terms of the predicted outcome of their representation Z (i.e., to be given the same predicted outcome

Y). To achieve individual fairness in Z, another adversary h_2: $\mathbb{R}^{d'} \rightarrow \{0, 1\}$ is included. Specifically, for each representation $z_i = f(x_i) \in D$, h_2 predicts an outcome $h_2(z_i) \in \{0, 1\}$ (of attribute Y) such that, for another representation $z_j = f(x_j)$, if x_i and x_j are similar (e.g., x_j is a nearest neighbor of x_i), the predicted outcome of z_j should be *distinguishable* with the predicted outcome of z_i, i.e., $h_2(z_j) \neq h_2(z_i)$. We formalize the individual (fairness) loss function on $h_2(f(X))$, denoted by $L_i(h_2(f(X)))$ (written as L_i for simplicity), as follows to capture the above objective, where a conceptual notation $h_2(Z) = h_2(f(X))$ is also used here to denote the process of generating all $h_2(z_i)$ for $z_i \in Z$.

$$L_i = L_i(h_2(f(X))) = \frac{\sum_{x_i \in D} \sum_{x_j \in \mathcal{N}_D^k(x_i)} |h_2(f(x_i)) - h_2(f(x_j))|}{k \cdot N} \tag{4}$$

When L_i is larger, $h_2(f(x_i)) \neq h_2(f(x_j))$ holds for more pairs of similar data points x_i and x_j in D. Thus, the goal of adversary h_2 is to *maximize* L_i so that h_2 is more capable of distinguishing similar data points.

To find the k nearest neighbors of a data point in D, a suitable similarity metric is needed. In this work, we choose the Euclidean distance (a commonly applied metric) on all features X as the similarity metric, but not the representations $f(X)$ for distance computation. This is to ensure that we find the data points that are "really" similar to their original features. Note that another similarity metric (that could be more suitable for a specific dataset) also works, which only influences the result of finding the nearest neighbors.

Total Loss. The total loss function $L(f, g, h_1, h_2)$ is formalized to be the weighted sum of the classification loss function, group loss function and individual loss function based on three coefficients α, β and δ, respectively.

$$L(f, g, h_1, h_2) = \alpha \cdot L_c + \beta \cdot L_g + \delta \cdot L_i \tag{5}$$

The coefficients α, β and δ provide a trade-off among accuracy, group fairness and individual fairness. We train our model with a min-max optimization: $\min_{f,g} \max_{h_1,h_2} \mathbb{E}_{X,A,Y}[L(f, g, h_1, h_2)]$ following adversarial learning [15].

Training Algorithm. We train our model in a number of epochs. In each epoch, we first sample a mini-batch D' from the dataset D. Next, we do the training for this epoch in 3 steps. In Step 1 and Step 2, we freeze the parameters of f and g, and then, we train the group adversary h_1 and individual adversary h_2, respectively, such that their objective functions are maximized. Finally, in Step 3, f and g are trained such that the total loss function $L(f, g, h_1, h_2)$ on D' is minimized. In this way, the group fairness and individual fairness can both be improved in the generated representation Z, and meanwhile the accuracy of classifier g, which is encoded in the total loss function, is also improved.

Although it is not theoretical guaranteed that the adversarial learning will always converge, several heuristics that we apply could encourage its convergence practically including training sufficient epochs and using mini-batches [17, 18]. In our algorithm, we aim at optimizing the group fairness and the individual fairness, and finally, our results in Section 5 show the balanced trade-off between the two targets (e.g., 30% improvement of group fairness under the same level of individual fairness). This verifies the practical convergence of our algorithm.

4.3 Theoretical Properties of Loss Functions

We give the theoretical properties to show the effectiveness of using our loss functions to ensure fairness. First, we show that the optimal value of L_g can upper-bound the demographic parity gap of any classifier trained on representation Z. In the supplementary material [16], we provide the proofs.

Lemma 1. *For a group adversary h_1, the optimal value of $L_g(h_1(Z), A)$ (denoted by $L_g(h_1^*(Z), A)$) is at least the demographic parity gap of any classifier η on representation Z, i.e., $L_g(h_1^*(Z), A) \geq \Delta DP_Z(\eta)$.*

In Lemma 1, we connect $L_g(h_1(Z), A)$ with $\Delta DP_Z(\eta)$ (i.e., the performance of Z), and thus we can obtain the worst $\Delta DP_Z(\eta)$ performance of any classifier trained on Z given the optimal group adversary h_1^*. This shows the effectiveness of using $L_g(h_1(f(X)), A)$ as the group loss function.

Analogously, we want to show the effectiveness of the individual loss function $L_i(h_2(Z))$. We consider the yNN "variant" of a classifier η trained on representation Z, denoted by $\Delta yNN'_{Z,k}(\eta)$, which is the same as the yNN metric except that the k-NN of any sample $z_i (= f(x_i))$ for $z_i \in Z$ are defined based on the original dataset D (namely, $\mathcal{N}_Z^k(z_i) = \{f(x_j) | x_j \in \mathcal{N}_D^k(x_i)\}$). This is to ensure that the measurement is based on the "real" similarity relationships of the data points. Lemma 2 shows that, for any classifier η trained on Z, $\Delta yNN'_{Z,k}(\eta)$ is lower-bounded by a value related to the optimal value of $L_i(h_2(Z))$.

Lemma 2. *For an individual adversary h_2 and any classifier η on representation Z, $\Delta yNN'_{Z,k}(\eta) \geq 1 - L_i(h_2^*(Z))$, where $L_i(h_2^*(Z))$ denotes the optimal value of $L_i(h_2(Z))$.*

In Lemma 2, we can also obtain the worst $\Delta yNN'_{Z,k}(\eta)$ performance given the optimal individual adversary h_2^*, showing that our individual loss L_i is effective.

4.4 Optimization with Focal Loss

To this end, we have formed our GIFair structure. However, we notice that the ranges of the three losses in Equation 5 have large differences (e.g., the value of L_i is much smaller than the other two losses). Since our target is to minimize the total loss, the loss with a smaller value receives less attention.

To solve this issue, we exploit the focal loss function [19] to alleviate the imbalance among the three losses. Consider an item with two possible outcomes 1 and 0. Let p be the estimated probability with outcome 1. We define a variable p_t to be p if the *true* outcome of this item is 1 and to be $1 - p$ otherwise. The formulation of *<u>F</u>ocal <u>L</u>oss function* is $FL(p_t) = -(1 - p_t)^\gamma \cdot \log(p_t)$, where $\gamma \geq 0$ is a focusing parameter and $(1-p_t)^\gamma$ is regarded as a *weight* term. We notice that if the value of p_t is high, its weight $(1 - p_t)^\gamma$ will be low. Thus, less (resp. more) weight is given to an item with higher (resp. lower) p_t value. Based on this idea, we re-design our total loss function by adjusting the weights of the three terms:

$$L(f, g, h_1, h_2) = (1 - L_c)^\gamma \cdot L_c + (1 - L_g)^\gamma \cdot L_g + (1 - L_i)^\gamma \cdot L_i \qquad (6)$$

Table 1: Statistics of Datasets

Dataset	**Train/Test**	**Protected Attribute** ($A = 1/0$)	$P(A = 1)$	$P(Y = 0)$
COMPAS	4,321/1,851	*race* (African-Americans/other races)	0.34	0.54
Adult	30,162/15,060	*gender* (females/males)	0.33	0.75
German	700/300	*age* (the aged/the young)	0.27	0.7

In this equation, if the value of one loss is small (resp. large), its weight is large (resp. small). In this way, we can balance the values of the three losses with their weights. Each loss could receive similar attention during training.

5 Experiments and Analysis

In this section, we conducted extensive experiments to evaluate the effectiveness of GIFair. We used three common real datasets: COMPAS, Adult and German. Table 1 lists the statistics. **COMPAS** [20] is used to predict whether a criminal defendant will recidivate ($Y = 1$) or not ($Y = 0$). **Adult** [21] is used to predict each person's income ($Y = 1$ if income $>$ 50K/y, and $Y = 0$ otherwise). **German** [22] classifies each individual as good ($Y = 0$) or bad ($Y = 1$) credit risks.

We selected LAFTR [13], LFR [7], iFair [12] and DualFair [10] as baselines. We also include UNFAIR, which is a normal classification algorithm that does not consider fairness. If the original loss function (i.e., Equation 5) is used, our algorithm is denoted as GIFair, while GIFair-focal denotes our algorithm on the focal loss function (i.e., Equation 6).We implemented all algorithms in Python.

We focus on the classification accuracy, group fairness and individual fairness. (1) For accuracy, we use *accuracy* (denoting ACC) which is defined to be 1 minus the average difference between the outcome and the predicted outcome of all data points, and *F-1 score* (denoting $F1$) which is defined to be the harmonic mean of the precision and the recall of a classifier. (2) For group fairness, we adopt the two metrics as introduced in Section 3, namely *demographic parity gap* (denoting ΔDP) and *equalized odds distance* (denoting ΔEO). (3) For individual fairness, we use *yNN*, denoted by ΔyNN (introduced in Equation 2).

We varied β and δ in GIFair from 0.1 to 20, while α is fixed to 1. For baselines, we also changed their coefficients from 0.1 to 20. For GIFair-focal, we varied γ from 0.05 to 5. By default, we set k to 10 when computing the k-nearest neighbors for yNN according to [12]. For each coefficient setting and each model, we trained it 5 times (using different random seeds) and obtained the mean performance on the test datasets. The implementation details of algorithms are included into the supplementary materials [16]. In the following, we show the experimental results.

Overall Comparison. Due to lack of space, we show the overall comparison of our GIFair algorithm with all baselines for the best value achieved for each measurement in [16]. GIFair outperforms all the baselines on most metrics.

Trade-off Studies. We studied the trade-off between any two terms from accuracy, group fairness and individual fairness. We compared with the baselines that also study the trade-off. To show which algorithm performs better under multi-metrics, we plotted the Pareto front curves (widely used in existing trade-off studies [12, 13], which only shows the dominating points of multi-metrics

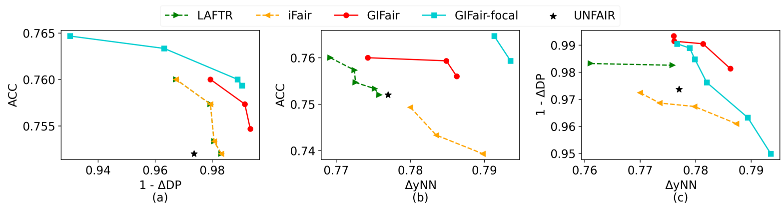

Fig. 2: Trade-off Curves on Dataset German

for better illustration). We also include baseline UNFAIR without weights for trading-off (thus shown as a star mark). Since the group fairness metrics are favored with smaller values, we plot 1 minus the group fairness metric, so that for each figure, the right-top points (high values along each axis) are preferable. We show the results on dataset German, while we obtain similar results for the other two datasets, which are reported in our supplementary material [16].

Accuracy and Group Fairness. Figure 2(a) shows the trade-off between accuracy and group fairness, with the default metric ACC and ΔDP, respectively. Compared with baselines, both GIFair and GIFair-focal have superior trading-off ability by reaching the most upper-right location. More closely, at the same level of accuracy ($ACC \approx 0.76$), the best ΔDP that baselines could achieve is at least 0.03 (i.e., $1 - \Delta DP < 0.97$), while the ΔDP values of our GIFair and GIFair-focal are around 0.02 and 0.01, improving the best baseline by 33% and 67%, respectively. For the same level of group fairness achieved (e.g., $\Delta DP \approx 0.02$), our GIFair and GIFair-focal obtain slightly better accuracy. The above indicates our better reconciliation between group fairness and accuracy compared with baselines, because we use an effective group fairness target, which ensures group fairness more easily without sacrificing accuracy too much. GIFair-focal could reach the highest ACC of around 0.765 but at a cost of sacrificing group fairness.

Accuracy and Individual Fairness. Figures 2(b) shows the trade-off between accuracy and individual fairness. GIFair and GIFair-focal still obtain the best trade-off. When ACC is fixed to around 0.76, the baseline with the best individual fairness has around 0.772 ΔyNN, while the ΔyNN of GIFair-focal reaches 0.792 with 2.6% improvement. Moreover, the baseline iFair could also obtain high ΔyNN of around 0.79 but with its ACC below 0.74, while our GIFair-focal keeps ACC above 0.76, which improves iFair by more than 3%. This similarly indicates that our algorithms better reconcile individual fairness and accuracy than iFair even though iFair has the same individual fairness target, because using adversarial learning could achieve the reconciliation more effectively.

Group Fairness and Individual Fairness. Our algorithms also obtain superior trade-off between the two types of fairness as shown in Figure 2(c). GIFair-focal achieves the highest ΔyNN (0.794), since it uses the focal loss function to effectively give larger weight to individual fairness while down-weigh group fairness. GIFair could also obtain good individual fairness (e.g., $\Delta yNN = 0.786$), while its group fairness is only slightly downgraded (with $\Delta DP = 0.02$).

Ablation Studies. We conducted ablation studies for the two adversaries in GIFair with the following variants. (1) GIFair without group adversary h_1 (i.e.,

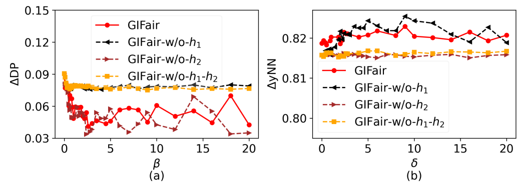

Fig. 3: Ablation Studies of GIFair on Dataset German

GIFair-w/o-h_1), by skipping Step 1 of training h_1. (2) GIFair without individual adversary h_2 (i.e., GIFair-w/o-h_2), by skipping Step 2 of training h_2. (3) GIFair without h_1 and h_2 (i.e., GIFair-w/o-h_1-h_2), by skipping both Step 1 and Step 2.

Figure 3 (a) and (b) illustrate the ablation study results on dataset German. Without group adversary h_1, GIFair-w/o-h_1 has much larger ΔDP (i.e., worse group fairness) than the original GIFair. This verifies the effectiveness of improving group fairness using the group adversary. Similarly, GIFair has larger yNN than GIFair-w/o-h_2, indicating that the individual adversary h_2 could effectively improve individual fairness. Without both adversaries, GIFair-w/o-h_1-h_2 obtains bad performance for both group and individual fairness.

Case Studies. We conducted case studies for the classification results regarding group and individual fairness. When only individual fairness is optimized (i.e., setting β to 0) for dataset COMPAS, we observe a representative result where 47% of the African-American group will recidivate, while this proportion for the other races is only 29%. When both group and individual fairness are optimized (i.e., setting all parameters to 1), the recidivation proportions among African-Americans and other races are predicted to be 40% and 38%, respectively, which is much fairer. Moreover, there exist some pairs of similar defendants who only have 1 day difference on the days between screening and arrest and have the same value for all other attributes. When only group fairness is optimized (i.e., setting δ to 0), we found that the number of these pairs of similar defendants that obtain different prediction results is 14. This number improves to only 1 when both group and individual fairness are optimized.

6 Conclusion

In this paper, we propose an adversarial learning structure, GIFair, with two adversaries for group fairness and individual fairness, respectively. With a designed training algorithm, GIFair can reconcile utility with group and individual fairness during generating a representation on the original dataset. We also propose a focal loss function that can better balance all the goals in GIFair. In our experiments on 3 real datasets, GIFair outperforms baselines with better fairness and higher accuracy. For future work, we would like to achieve a holistic optimization for utility and multiple fairness goals at the same time, and explore the problem on intersectional or unknown group.

Acknowledgements We greatly thank Zheng Zhang for his contribution on this paper.

References

1. M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa*ir: A fair top-k ranking algorithm," in *CIKM*, 2017, pp. 1569–1578.
2. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NeurIPS*, 2016, pp. 3323–3331.
3. F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," in *KAIS*, vol. 33, 2011, pp. 1–33.
4. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ITCS*, 2012, pp. 214–226.
5. M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.
6. B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Interventional fairness: Causal database repair for algorithmic fairness," in *SIGMOD*, 2019, pp. 793–810.
7. R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, vol. 28, no. 3, 2013, pp. 325–333.
8. H. Edwards and A. Storkey, "Censoring representations with an adversary," in *ICLR*, 2016.
9. H. Zhao, A. Coston, T. Adel, and G. J. Gordon, "Conditional learning of fair representations," in *ICLR*, 2020.
10. S. Han, S. Lee, F. Wu, S. Kim, C. Wu, X. Wang, X. Xie, and M. Cha, "Dualfair: Fair representation learning at both group and individual levels via contrastive self-supervision," *arXiv preprint arXiv:2303.08403*, 2023.
11. R. Binns, "On the apparent conflict between individual and group fairness," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020.
12. P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *ICDE*, 2019, pp. 1334–1345.
13. D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *ICML*, vol. 80, 2018, pp. 3384–3393.
14. D. Kim, K. Kim, I. Kong, I. Ohn, and Y. Kim, "Learning fair representation with a parametric integral probability metric," *arXiv preprint arXiv:2202.02943*, 2022.
15. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.
16. H. Liu and R. C.-W. Wong, "Adversarial learning of group and individual fair representations (supplementary material)," 2024. [Online]. Available: https://github.com/satansin/GIFair
17. D. Saxena and J. Cao, "Generative adversarial networks (gans) challenges, solutions, and future directions," *ACM Computing Surveys*, vol. 54, no. 3, 2021.
18. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.
19. T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *ICCV*, pp. 2999–3007, 2017.
20. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: Risk assessments in criminal sentencing," ProPublica, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
21. B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.
22. H. Hofmann, "Statlog (German Credit Data)," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C5NC77.