Adversarial Learning of Group and Individual Fair Representations (Technical Report)

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 fairness. Compared with most relevant studies that only focus on group fairness without individual fairness, GIFair makes sure that the classifiers trained on the generated representation satisfy both individual fairness and group fairness. A theoretical proof is provided to show that except in highly constrained special cases, group fairness and individual fairness cannot be satisfied simultaneously, and thus, we need to trade group fairness off against individual fairness in addition to considering the utility of classifiers. Experiments conducted on three real datasets show that GIFair can achieve a better utility-fairness trade-off compared with existing models.

ACM Reference Format:

. 2023. Adversarial Learning of Group and Individual Fair Representations (Technical Report). In *Proceedings of SIGKDD (Conference acronym 'XX)*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Machine learning models are widely adopted to help make decisions nowadays. Nearly all classification tasks only aim to achieve high utility, but the fairness of machine learning models is often overlooked by researchers. It has been shown in many studies [4, 23] that if the historical datasets are biased against some groups of people, machine learning models trained on these biased datasets to pursue high accuracy will create discrimination in the decision making process. For example, due to the racism in the US criminal justice system, the rates of arrest and conviction of African-Americans are extremely higher than other races. If we use these criminal records to predict recidivism, it will be biased against African-Americans [2].

Discrimination, which means a person or a particular group is treated differently (especially in a worse way) due to his/her race, gender, or sexuality, can be reflected in many aspects of society. 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 example, a bank officer evaluates the credit levels of applicants based on the information of applicants (e.g., age, gender and credit history) to decide whether to approve their loans. It is less likely that applications from women are approved [31]. In this case, women are the protected group. Unfairness is reflected in not only gender but also race. In 2017, Sara et al. [5] studied 3453 American adults and found that African-Americans have a significant portion reporting discrimination. These examples show the existence of systemic discrimination and we need to address racism or sexism more actively. Motivated by this, we want to propose a fair classification model to help to alleviate discrimination in decision-making systems.

To check the fairness of different classification models, many fairness notions have been proposed and most of them can be divided into *group fairness* [13, 15] and *individual fairness* [8]. Group fairness requires classifiers to treat different groups defined by protected attributes equally. One popular notion of group fairness called *demographic parity* requires that classification is *independent* of the protected attributes. On the other hand, individual fairness requires *similar* individuals should be treated *similarly* by classifiers.

Based on these fairness notions, many approaches have been proposed to solve the fair classification problem. Some methods process the historical datasets to mitigate discrimination directly by modifying the original outcomes [14] or the attributes of data [15]. Some methods achieve fairness during training by setting fairness as a regularizer [17, 18, 32] or hard constraints [6]. Representation learning [9, 33] is another common approach. The idea of representation learning is to transform original datasets into new representations and obfuscate the information about the protected attributes in the representations. Then, the representations of different groups are similar and different groups will be treated similarly by any classifier. In this way, this method satisfies group fairness.

However, most existing studies only focus on group fairness but do not address the problem of reconciling group fairness and individual fairness *at the same time*. In addition to group fairness, individual fairness is also a very important aspect of fairness. Only satisfying group fairness in machine learning models may harm individual fairness, which could create discrimination. For example, according to [32], in hiring decision, some unqualified people in the protected group (e.g., females) are interviewed deliberately so that demographic parity is satisfied among all candidates interviewed, which is, in fact, biased against the unprotected group and is contrary to the requirement of individual fairness. Individual fairness can alleviate this kind of discrimination by ensuring that any two individuals who are similar in terms of attributes/background (e.g., similar academic experience) are treated similarly.

There are only a few studies [3, 10, 27, 32] that consider both individual and group fairness in their design goals. Our most closely-related work is LFR [32]. LFR addresses the classifier accuracy, group fairness and individual fairness for classification by defining

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

a combined loss function that is a weighted sum of the three terms. Then, it optimizes the combined loss function during learning a new representation by mapping items in the original dataset to a set of *prototypes* probabilistically where a prototype could be viewed as a cluster in LFR. One limitation of LFR is that the three terms in the objective function are trained at the same time, and thus, the three terms may not be reconciled well *at the same time*. Besides, the design of loss functions in LFR enforces fairness *indirectly*, so the fairness performance of learned representation is not guaranteed. Some studies address a different machine learning task known as *fair ranking* [10, 27], which aims to rank the individuals without bias. In [10], an objective function of individual unfairness is minimized while some hard constraint for group fairness is satisfied for ranking, and thus there is no trade-off or reconciliation between group and individual fairness. [27] targets the trade-off between utility and each of the two fairness notions separately, but does not study the trade-off between individual and group fairness. Another work [3] attempts to achieve the compatibility of individual and group fairness by a data-driven model based on the received unfairness complaints. However, it is not easy to obtain the data about unfairness complaints which prevents this model from practical uses. In [3], the authors demonstrate the results in artificial datasets only.

We mainly focus on reconciling accuracy and two kinds of fairness (i.e., group fairness and individual fairness) in this work. To solve this problem, we propose an approach called **GIFair** (for group fair and individual fair representations) to transform the original dataset into a *fair representation*. Different from most previous studies that only focus on one kind of fairness, we study how to achieve group fairness and individual fairness at the same time in the learned representation. To achieve this goal, we use two adversaries, one for group fairness and the other for individual fairness, instead of using only one adversary in the related studies. To reconcile the two adversaries such that accuracy, individual fairness and group fairness are achieved at the same time in the generated representation, we propose a well-designed training algorithm to reconcile all concepts (i.e., accuracy, individual fairness and group fairness) in our structure. Different from the traditional adversarial learning studies that only need to consider accuracy and group fairness, our proposed training algorithm can handle such 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.

We also give theoretical proof showing that group fairness and individual fairness cannot be satisfied simultaneously in most cases, so we need to do a trade-off between group fairness and individual fairness. To the best of our knowledge, there is no prior work on this problem. The relation between the two kinds of fairness is not well studied in previous studies [3, 10, 27, 32]. [32] empirically only *assumes* that the two kinds of fairness are opposing without giving any *theoretical analysis*. We also 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 individual fairness performance on dataset Adult.

The contributions of our work are summarized as follows.

- We design a novel structure of adversarial representation learning with two adversaries, one for the individual fairness and the other for the group fairness.
- We design a training algorithm that can well reconcile the two adversaries in our structure. Ablation analysis is conducted to show its superiority.
- We *theoretically* show the incompatibility of group fairness and individual fairness, which is the first result in the literature.
- The experiments conducted on three real datasets empirically 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 of the fair classification problem. 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 the following three categories, namely *pre-processing*, *in-processing* and *post-processing*. *Pre-processing approaches* [15, 26, 28] directly modify datasets to remove discrimination before using normal methods to do classification. [15] pre-processes the data to remove discrimination by suppressing the protected attribute, resampling the data and so on. [26] uses a causal approach to remove discrimination in datasets and provides fairness guarantees about all classifiers trained on the dataset after processed. [28] alleviates bias by removing some data points during training. *In-processing approaches* [6, 17, 27, 32] modify the classifier to improve its fairness performance while maintaining the utility of classification during training. The common trick is using regularizers in the loss function to balance two goals: maximizing the accuracy of classifiers and minimizing the discrimination in prediction results. For example, [17] probabilistically maps all items of the dataset into a low-rank representation that reconciles individual fairness and the utility of classifiers. Methods [6, 27] enforce fairness during training by modeling fairness as hard constraints. *Post-processing approaches*, e.g., [13], directly change the predicted outcomes of the learned predictors.

Among all those approaches, we review some branches that are mostly related to this work.

Learning Fair Representations. Recently, *fair representation learning* attracts great attention in fair machine learning, with LFR [32] (introduced in Section 1) as the first work in this line. Fair representation learning is to learn a debiased representation of the original dataset so that the downstream tasks on this representation could satisfy fairness requirements. Many approaches have been proposed to learn fair representations. [12] proposes a graph-based regularization approach under a group fairness constraint to decrease the dependence between the protected attribute and the representation. [24] adopts contrastive learning to learn the disentangled invariant representation such that the representation space is separated into two parts, one of which is unrelated to the protected attribute. [21] proposes distance covariance between the representation and the protected attribute as a new dependence measurement.

Among those approaches, adversarial representation learning has been broadly explored. ALFR [9] provides a framework to mitigate discrimination by learning representations that minimize the performance of the adversary that is trained to predict the protected attribute of the representation. LAFTR [22] follows this framework to explore adversarial learning as a method of obtaining a representation to mitigate unfair prediction outcomes. LAFTR proves that the learned representation can lead to group fair prediction. DCFR [30] defines a new fairness metric called *conditional fairness* by conditioning on the pre-decided fairness variables and proposes an adversarial representation learning algorithm to achieve conditional fairness. IPM [16] proposes the integral probability metric adopted in an adversary such that a good theoretical guarantee on group fairness is obtained.

Our method GIFair follows the idea of adversarial representation learning. However, instead of only focusing on group fairness in all the above studies of learning fair representations (except LFR), we consider how to achieve individual fair prediction at the same time.

Trade-off between Accuracy and Fairness. There are also some previous studies focusing on the trade-off between accuracy and fairness. [29] aims to improve the trade-off between group fairness and accuracy, but this study focuses on solving the problem in the multi-task setting. [25] presents a theoretical analysis on the difficulty of obtaining group fairness and accuracy simultaneously, and shows some restricted conditions such that group fairness and accuracy could be compatible. [19] adapts a Siamese network approach to achieve the trade-off between accuracy and individual fairness. [27] targets the trade-off between accuracy and each of the two fairness notions, individual and group fairness, separately, and this work addresses the ranking problem which is different from our focus (i.e., classification). Nevertheless, none of the above studies involve the trade-off between group fairness and individual fairness, which is addressed in this work.

3 PRELIMINARIES

3.1 Notations

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, each of which is a scalar attribute. Thus, x_i is represented by a vector in the d-dimensional space, i.e., $x_i \in \mathbb{R}^d$. Data point x_i is also associated with a (categorical) outcome attribute Y for classification and a (categorical) protected attribute A representing the group membership (e.g., gender). Dataset D can thus be divided into different groups (e.g., female group and male group).

In this work, we first focus on the case where the outcome attribute and the protected attribute are both binary, and thus we assume that the domains of both Y and A are $\{0, 1\}$. We discuss the multi-outcome and multi-group case in Section 4.5. We assume that values 1 and 0 represent the protected group (e.g., female group) and the unprotected group (e.g., male group), respectively. We thus denote D_1 and D_0 to be the subsets of D containing all data points in the protected group and the 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 criterion is satisfied. In the next section, we introduce the fairness notions used to form our fairness criterion.

3.2 Fairness Notions

Many fairness notions were proposed in the recent literature. For group fairness, two popular notions are *demographic parity* [8] and *equalized odds* [13]. Demographic parity requires that the success rates (i.e., the rates of positively predicted outcomes) of all protected groups and non-protected groups are equal. Equalized odds requires that the false positive and true positive rates should also be equal among different groups. However, the above fairness notions may not be exactly satisfied by classifiers in most cases. Thus, following a common approach, we use *demographic parity gap* to measure how well a classifier satisfies group fairness. Specifically, given a classifier η and dataset D, the demographic parity gap of η for D, denoted by $\Delta DP_D(\eta)$, is defined as follows.

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

Here, we use $E_{D_1}(\eta)$ (resp. $E_{D_0}(\eta)$) to denote the proportion of data points whose outcomes are predicted as 1 in D_1 (resp. D_0). Clearly, if the difference between $E_{D_1}(\eta)$ and $E_{D_0}(\eta)$ (i.e., $\Delta DP_D(\eta)$) is smaller, the proportion of members in each group predicted as 1 among all members in the group is more balanced, which indicates better group fairness for both groups.

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 [8]. Consider a data point x_i. Let $k\text{-}NN_D(x_i)$ denote the set of k nearest neighbors of x_i in D, where k is a positive integer. Note that $k\text{-}NN_D(x_i)$ 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* [32], 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 as follows.

$$\Delta yNN_{D,k}(\eta) = 1 - \frac{\sum_{x_i \in D} \sum_{x_j \in k\text{-}NN_D(x_i)} |\eta(x_i) - \eta(x_j)|}{k \cdot N} \quad (2)$$

This metric captures the average difference between the predicted outcome of a data point x_i and the predicted outcome of a nearest neighbor of x_i. This difference is 0 if the two similar data points has the same predicted outcome and 1 otherwise. Thus, according to Equation 2, when $\Delta yNN_{D,k}(\eta)$ is larger, we could achieve a better individual fairness.

In particular, when $\Delta yNN_{D,k}(\eta) = 1$, η is said to satisfy a special individual fairness requirement called the *yNN condition* for dataset D. That is, a classifier η satisfies the yNN condition for D if the predicted outcome of any data point x_i in D is the same as the predicted outcomes of all the k nearest neighbors of x_i.

3.3 Generative Adversarial Network

Generative adversarial network (GAN) is an adversarial network [11] consisting of two components, namely a *generator* G and a *discriminator* C. The generator G aims at deceiving the discriminator C by constructing synthetic data from a prior distribution P_z on a noise variable z to match the real data distribution P_{data}.

The discriminator C is a binary classifier that aims at distinguishing whether the data comes from real data distribution P_{data} or synthetic data $G(z)$ constructed by generator G.

Both components improve their ability through learning. That is, G is trained to generate $G(z)$ that cannot be distinguished from the real data by C, and C is trained to identify the outcome of $G(z)$ more accurately. Then, the learning of GAN is formalized as a min-max optimization $\min_G \max_C V(G,C)$, where $V(G,C)$ is a total loss defined as follows.

$$V(G,C) = \mathbb{E}_{x\sim P_{data}}[\log(C(x))] + \mathbb{E}_{z\sim P_z}[1-\log(C(G(z))] \quad (3)$$

where discriminator C seeks to maximize $V(G,C)$ but generator G seeks to minimize $V(G,C)$.

In our work, we design a novel adversarial network with two different adversaries to obtain both group fairness and individual fairness, and meanwhile the prediction accuracy is also our design goal.

4 METHODOLOGY

4.1 Problem Statement

In this work, we follow adversarial representation learning to tackle the fair classification problem. Specifically, our fair classification problem 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. Specifically, a classifier η is fair in terms of group fairness, if a smaller demographic parity gap of η for D (i.e., $\Delta DP_\eta(D)$) is obtained, and η is fair in terms of individual fairness, if a larger yNN of η for D and k (i.e., $\Delta yNN_{D,k}(\eta)$) is obtained.

It is worth mentioning that the group fairness and individual fairness could not be satisfied simultaneously in most cases, which will be further elaborated in Section 4.3. Thus, we set our optimization goal of classifier η such that a balance can be obtained among accuracy, group fairness and individual fairness.

4.2 Model

First proposed by [9], plenty of existing work follows 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. To achieve that, an *adversary* using a *discriminator* is set up to identify the group of the generated representation Z. However, this framework so far only addresses group fairness. It remains unsolved how to accommodate individual fairness into this framework and how to obtain a reconciliation between different fairness targets together with classification accuracy.

With this motivation, we propose our model called **GIFair** (**G**roup **I**ndividual **Fair**). As illustrated in Fig. 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 the representation Z, seeks to preserve the prediction accuracy compared to making prediction from the original features X. In addition, GIFair aims at achieving group fairness by the group adversary h_1 and individual fairness

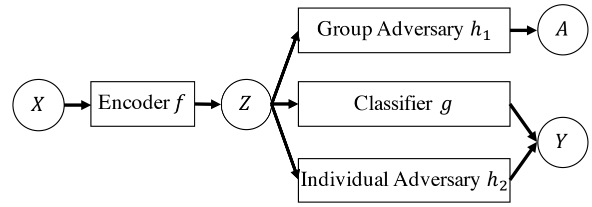

Figure 1: Structure of GIFair.

by the individual adversary h_2. We will introduce the details of all components and how they interact with each other in the following.

Encoder. An encoder $f\colon \mathbb{R}^d \to \mathbb{R}^{d'}$ maps a data point x_i into a vector in the d'-dimensional space, denoted by $z_i = f(x_i)$, which is called the representation of x_i. The representation Z of the original dataset is formed by the representations of all data points in D, namely, $Z = \{f(x_i)|x_i \in D\}$. In this way, the encoder f reconstructs the origin features X into the representation Z, and we thus use $Z = f(X)$ to conceptualize this re-construction process.

Classifier. While re-constructing X to Z with encoder f, the *utility* of X may be lost. Losing utility means that prediction with representation Z is not as accurate as prediction with the original features X concerning the outcome attribute Y. As such, we use a classifier $g\colon \mathbb{R}^{d'} \to \{0,1\}$ to predict the outcome $g(z_i)$ of each representation z_i in Z. Conceptually, let $g(Z)$ denote the prediction process of all representations in Z (note that $g(Z) = g(f(X))$). To preserve utility, our goal is to achieve accurate prediction of g concerning Y. We thus minimize a suitable classification loss function between $g(f(X))$ and Y in a dataset D, denoted by $L_{cls}(g(f(X)),Y)_D$ (which is selected to be cross-entropy in this work).

Group Adversary. To ensure that the generated representation Z achieves group fairness, the group adversary $h_1\colon \mathbb{R}^{d'} \to \{0,1\}$ is included in GIFair. 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. Note that encoding x_i to z_i does not alter the group membership of x_i, and thus the group of z_i is still defined by the protected attribute A of x_i. Conceptually, we use $h_1(Z) = h_1(f(X))$ to denote the process of predicting the group of all representations in Z for h_1. The objective of adversary h_1 is to best *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, we form the group fairness loss function on $h_1(f(X))$ and A in a dataset D, denoted by $L_{grp}(h_1(f(X)),A)_D$, as follows.

$$\begin{aligned} L_{grp}(h_1(f(X)),A)_D &= |F_{D_0\to 1}(h_1) - F_{D_1\to 1}(h_1)| \\ &= |\frac{\sum_{x_i\in D_0} h_1(f(x_i))}{|D_0|} - \frac{\sum_{x_j\in D_1} h_1(f(x_j))}{|D_1|}| \end{aligned} \quad (4)$$

Here, we use $F_{D_0\to 1}(h_1)$ (resp. $F_{D_1\to 1}(h_1)$) to denote the proportion of representations whose predicted group label (by h_1) is 1 among all representations whose group is originally 0 (resp. 1).

Intuitively, when the value of $L_{grp}(h_1(f(X)),A)_D$ is large, the difference between $F_{D_0\to 1}(h_1)$ and $F_{D_1\to 1}(h_1)$ is large, which falls to two cases. In the first case, $F_{D_0\to 1}(h_1)$ is small and $F_{D_0\to 1}(h_1)$

is large, indicating that the group label of only a few representations from D_0 are wrongly predicted, and the group label of most representations from D_1 are correctly predicted. In summary, correct predictions of a majority of the representations in Z are given by h_1. Similarly, in the second case where $F_{D_0 \to 1}(h_1)$ is large and $F_{D_0 \to 1}(h_1)$ is small, one can find that most representations in Z are wrongly predicted. It thus indicates that, for both cases, h_1 succeeds in making good differentiation of representations from different groups. Therefore, h_1 is trained to *maximize* $L_{grp}(h_1(f(X)), A)_D$.

Consider back the group fairness requirement. The generated representation Z should obfuscate the group information in A such that any classifier trained on Z will treat different groups equally. To achieve that, encoder f aims to fool h_1 by generating Z such that h_1 *cannot* easily differentiate representations in different groups. As such, Z is obtained by *minimizing* $L_{grp}(h_1(f(X)), A)_D$ in encoder f, which incurs a typical adversarial learning scheme.

One can observe that $L_{grp}(h_1(f(X)), A)_D$ has a similar formulation with demographic parity gap $\Delta DP_D(\eta)$ (defined in Equation 1) given a classifier η. We will show that our loss function $L_{grp}(h_1(f(X)), A)_D$ can upper-bound the demographic parity gap of any classifier η trained on representation Z in Section 4.3.

Individual Adversary. The third term in GIFair is individual fairness, which requires that individuals who are similar on their features X should be *indistinguishable* in terms of the predicted outcome of their generated representation Z. To achieve individual fairness in Z, another adversary h_2: $\mathbb{R}^{d'} \to \{0, 1\}$ is included. Specifically, for each representation $z_i = f(x_i) \in D$, h_2 predicts an outcome $h_2(z_i) \in \{0, 1\}$ 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 $h_2(f(X))$ in a dataset D, denoted by $L_{ind}(h_2(f(X)))_D$, 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 $z_i \in Z$.

$$L_{ind}(h_2(f(X)))_D = \frac{\sum\limits_{x_i \in D} \sum\limits_{x_j \in k\text{-}NN_D(x_i)} |h_2(f(x_i)) - h_2(f(x_j))|}{k \cdot N} \tag{5}$$

Clearly, when $L_{ind}(h_2(f(X)))_D$ is larger, $h_2(f(x_i)) \neq h_2(f(x_j))$ holds for more pairs of data points x_i and x_j in D where x_i and x_j are similar. Thus, the goal of adversary h_2 is to *maximize* $L_{ind}(h_2(f(X)))_D$ such that h_2 is more capable of distinguishing similar data points. However, to achieve individual fairness of representation Z, encoder f aims to make similar data points indistinguishable. Thus, f is trained such that $L_{ind}(h_2(f(X)))_D$ is *minimized.*

Note that $L_{ind}(h_2(f(X)))_D$ is similar to the formation of metric yNN for measuring individual fairness (defined in Equation 2), but we drop the "one minus" part from yNN to keep a maximization target of h_2 (to be consistent with the target of h_1). Moreover, a suitable similarity metric is needed to find the k nearest neighbors of a data point in D. In this work, we choose the Euclidean distance (a commonly applied metric) on all features X as the similarity metric.

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

$$\begin{aligned} L(f, g, h_1, h_2)_D = {} & \alpha \cdot L_{cls}(g(f(X)), Y)_D \\ & + \beta \cdot L_{grp}(h_1(f(X)), A)_D \\ & + \delta \cdot L_{ind}(h_2(f(X)))_D \end{aligned} \tag{6}$$

The coefficients α, β and δ provide a trade-off among accuracy, group fairness and individual fairness GIFair. We train our model with the following min-max optimization of our total loss function.

$$\min_{f,g} \max_{h_1,h_2} \mathbb{E}_{X,A,Y}[L(f, g, h_1, h_2)_D] \tag{7}$$

where the two adversaries h_1 and h_2 are trained separately to maximize the total loss. However, encoder f and classifier g are trained jointly to minimize total loss.

Training. The pseudo-code of our learning algorithm of GIFair is shown in Algorithm 1. Let θ_f, θ_g, θ_{h_1} and θ_{h_2} denote the trainable parameters of encoder f, classifier g, group adversary h_1 and individual adversary h_2, respectively. Our algorithm runs in e epochs, where e is the input number of epochs. At the beginning of each epoch, we sample a mini-batch D' from the dataset D where the size of D' is an input parameter m. Next, we do the training for this epoch in 3 steps. In Step 1 (Line 4-5) and Step 2 (Line 7-8), 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. This is done by ascending along the gradients of their objective functions on D', namely $L_{grp}(h_1(f(X)), A)_{D'}$ and $L_{ind}(h_2(f(X)))_{D'}$. Note that Steps 1 and 2 have no dependency, and thus the ordering between them could be reversed. After Step 1, the ability of h_1 distinguishing representations in different groups is improved. Similarly, the ability of h_2 predicting different outcomes between a representation and its original neighbors is improved after Step 2. Finally, in Step 3 (Line 10-11), f and g are trained such that the total loss function $L(f, g, h_1, h_2)_{D'}$ on D' is minimized, by descending along the gradients of $L(f, g, h_1, h_2)_{D'}$. 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.

4.3 Theoretical Analysis

In this section, we first show that group fairness and individual fairness cannot be both satisfied in most cases by showing that they can only be satisfied simultaneously in two highly constrained conditions in Theorem 4.1. This motivates our goal to obtain a trade-off between group fairness and individual fairness. Then, in Theorem 4.2, we show that the optimal value of $L_{grp}(h_1(f(X)), A)_D$ can upper-bound the demographic parity gap of any classifier η trained on representation Z (i.e., $\Delta DP_Z(\eta)$). This shows the effectiveness of using $L_{grp}(h_1(f(X)), A)_D$ as the group fairness loss function. In Section C, we provide the proofs of the theorems.

Trade-off between Group and Individual Fairness. *When are two kinds of fairness (i.e., demographic parity (for group fairness) and the yNN condition (for individual fairness)) simultaneously achieved?* To answer this question, we present the following Theorem 4.1, which leverages the concept of *k-NN cluster*. For any two data points $x_i, x_j \in D$, we connect them with an edge if $x_i \in k\text{-}NN_D(x_j)$ or $x_j \in k\text{-}NN_D(x_i)$. Then, the dataset D is modeled as an *undirected graph*. We define a k-NN cluster in D to be the set of all the data

Algorithm 1 GIFair

Input: Dataset D, Batch size m, Number of epochs e
Output: Encoder f, Classifier g, Group adversary h_1, Individual adversary h_2

```
for epoch from 1 to e do
    Randomly sample a mini-batch D' from D of size m
    ▷ Step 1 (Train h_1)
    Freeze h_2, f, g; Unfreeze h_1
    Optimize h_1 by ascending along gradient on D':
        ∇_{θ_{h_1}} L_grp(h_1(f(X)), A)_{D'}
    ▷ Step 2 (Train h_2)
    Freeze h_1, f, g; Unfreeze h_2
    Optimize h_2 by ascending along gradient on D':
        ∇_{θ_{h_2}} L_ind(h_2(f(X)))_{D'}
    ▷ Step 3 (Train f and g)
    Freeze h_1, h_2; Unfreeze f, g
    Optimize f and g by descending along gradient on D':
        ∇_{θ_f, θ_g} L(f, g, h_1, h_2)_{D'}
return f, g, h_1, h_2
```

points in a connected component of this graph. Given a k-NN cluster in D, says C, it is easy to observe that the nearest neighbor of any data point x_i in C is also in C, and thus $k\text{-}NN_D(x_j) \subseteq C, \forall x_j \in C$.

Theorem 4.1. *Consider a classifier η. η satisfies demographic parity and the yNN condition simultaneously, if and only if η satisfies either of the two conditions:*

1. $\{\eta : X \rightarrow \{0,1\} \mid \eta$ is a constant function$\}$;

2. (1) For any k-NN cluster C in D, $\eta(x_i) = \eta(x_j), \forall x_i, x_j \in C$ and (2) $Pr[\eta(x_i) = 1 | x_i \in D_1] = Pr[\eta(x_i) = 1 | x_i \in D_0]$.

However, the two conditions are both highly constrained special cases. In Condition 1, the constant function set (which gives the same prediction outcome to all data points) is a very restricted set [1]. The classifiers that satisfy Condition 2 can be seen as a variation of a constant function because the data points in the same k-NN cluster are given the same prediction result, and thus Condition 2 is also highly constrained. We can conclude that in most conditions, group fairness and individual fairness cannot be satisfied simultaneously. Besides, these conditions are not desirable especially when we want to design an accurate classifier. In summary, Theorem 4.1 shows the incompatibility between the two kinds of fairness, and hence we should find an optimal trade-off between them.

Effectiveness of Group Fairness Loss. Next, we theoretically analyze our proposed objective of group adversary in Theorem 4.2.

Theorem 4.2. *Consider a group adversary $h_1 : \mathbb{R}^{d'} \rightarrow \{0,1\}$. The optimal value of $L_{grp}(h_1(Z), A)_D$ (denoted by $L_{grp}(h_1^*(Z), A)_D$) is at least the demographic parity gap of any classifier $\eta : \mathbb{R}^{d'} \rightarrow \{0,1\}$ on representation Z, i.e., $L_{grp}(h_1^*(Z), A) \geq \Delta DP_Z(\eta)$.*

In Theorem 4.2, we connect the objective $L_{grp}(h_1(Z), A)_D$ 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 g trained on Z given the optimal adversary h_1^*. However, h_1^* may not be obtained during training but a sufficiently powerful h_1 of which $L_{grp}(h_1(Z), A)_D$ is close to $L_{grp}(h_1^*(Z), A)_D$ can be used as the upper bound of the $\Delta DP_Z(\eta)$ performance of any classifier g according to [22].

4.4 Optimized Weights with Focal Loss

During training by using the loss function of Equation (6), we note that the ranges of three losses are much different (e.g., the value of $L_{ind}(h_2(f(X)))_D$ 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. We could solve this problem by giving this loss a larger weight than the other two losses or we can propose a new loss function that can solve this problem.

Our new loss function uses the idea of focal loss function [20] originally used to address the class imbalance problem. Here, we want to use this idea to address the imbalance problem among these three losses. Consider an item with two possible outcomes, namely a positive outcome and a negative outcome. Let p be the estimated probability that this item has the positive outcome. 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 *focal loss function* is $FL(p_t) = -(1-p_t)^\gamma \cdot \log(p_t)$, where $\gamma \geq 0$ is a focusing parameter. Note that $(1-p_t)^\gamma$ is regarded as a *weight* term in this function. We notice that if an item with its true outcome equal to 1 is correctly classified, p_t is close to 1 so its weight $(1-p_t)^\gamma$ is close to 0. In this way, the focal loss function can down-weigh the weights assigned to all items with high p_t values. On the other hand, it could give (relatively) more weights assigned to items with low p_t values.

Based on this idea, we could re-design our total loss function by adjusting the weights of the three terms:

$$\begin{aligned} L(f,g,h_1,h_2)_D = {} & (1 - L_{cls}(g(f(X)),Y)_D)^\gamma \cdot L_{cls}(g(f(X)),Y)_D \\ & + (1 - L_{grp}(h_1(f(X)),A)_D)^\gamma \cdot L_{grp}(h_1(f(X)),A)_D \\ & + (1 - L_{ind}(h_2(f(X)))_D)^\gamma \cdot L_{ind}(h_2(f(X)))_D \end{aligned} \tag{8}$$

The weights given to the three losses are similar to the weight in the focal loss. That is, we use the value of each loss in the new weights. If the value of one loss is small (e.g., the loss of individual adversary), its weight is large. On the other hand, weights are small for large values of losses. In this way, we can balance the values of the three losses with their weights. Each loss could receive similar attention during training.

4.5 Handling Multi-outcome and Multi-group

In this section, we discuss how our GIFair model can be adapted to handle multiple values for the outcome attribute Y and multiple group values for the protected attribute A. We formalized our total loss function consisting of the three loss functions for accuracy (using cross-entropy), group fairness (Equation 4) and individual fairness (Equation 5). Now, we present how the three loss functions are modified to handle multi-outcome and multi-group, while the total loss remains the same weighted sum formation.

Firstly, for the classification loss $L_{cls}(g(f(X)),Y)_D$, since it uses the cross-entropy form, it is easily adapted to multi-outcome case.

Secondly, for the group fairness loss function, when the number of groups is more than 2, our intuition is to first consider a group fairness loss for every pair of groups (using a form similar to $L_{grp}(h_1(f(X)),A)_D$ defined on two groups), and then aggregate

Table 1: Statistics of Datasets

Dataset	Train/Test	$P(A=1)$	$P(Y=0)$
COMPAS	4,321/1,851	0.34	0.54
Adult	30,162/15,060	0.33	0.75
German	700/300	0.27	0.7

the losses for all pairs. Consider the multi-group domain of A to be $\{1, 2, ..., M\}$, where M is the total number of groups. The protected attribute of each data point x_i is thus an integer between 1 and M representing group membership of x_i. Let D_r denote the set of all data points in Group r, where $r \in [1, M]$. We form the multi-group fairness loss function, denoted by $L'_{grp}(h_1(f(X)), A)_D$, as follows.

$$L'_{grp}(h_1(f(X)), A)_D = \frac{1}{M^2} \sum_{r=1}^{M} \sum_{s=1}^{M} |F_{D_r \to r}(h_1) - F_{D_s \to r}(h_1)| \quad (9)$$

where $F_{D_r \to r}(h_1)$ (resp. $F_{D_s \to r}(h_1)$) denotes the proportion of representations whose predicted group label (by h_1) is r among all representations originally in Group r (resp. s). Intuitively, for every *ordered* pair of groups r and s, when the above multi-group fairness loss is large, we also have two cases, one with small $F_{D_r \to r}(h_1)$ and large $F_{D_s \to r}(h_1)$, the other with large $F_{D_r \to r}(h_1)$ and small $F_{D_s \to r}(h_1)$. For the first case, most representations from D_r are *not* predicted to be r while most representations from D_s are predicted to be r. This incurs that h_1 can well differentiate representations from D_r and D_s (note that the second case also leads to this conclusion). Since $|F_{D_r \to r}(h_1) - F_{D_s \to r}(h_1)|$ only captures the differentiating ability of h_1 concerning group D_r, D_s and predicted label r, we aggregate the result for all ordered pairs of groups to form our multi-group fairness loss function $L'_{grp}(h_1(f(X)), A)_D$. Identically, adversary h_1 will be trained to maximize this loss function as its objective.

Thirdly, for the individual fairness loss function, a simple modification is performed on $L_{ind}(h_2(f(X)))_D$ to form a multi-outcome individual fairness loss function $L'_{ind}(h_2(f(X)))_D$. Specifically,

$$L'_{ind}(h_2(f(X)))_D = \frac{\sum\limits_{x_i \in D} \sum\limits_{x_j \in k\text{-}NN_D(x_i)} \mathcal{F}(h_2(f(x_i)) - h_2(f(x_j)))}{k \cdot N} \quad (10)$$

where $\mathcal{F}(x)$ returns 0 for $x = 0$ and returns 1 for any non-zero x. Clearly, when h_2 predicts a multi-value outcome of the representation of a data point x_i, adversary h_2 only needs to give a different outcome for a nearest neighbor x_j of x_i, i.e., $h_2(f(x_i)) \neq h_2(f(x_j))$. Thus, to remove the influence of concrete outcome values, as long as $h_2(f(x_i)) \neq h_2(f(x_j))$, value 1 will be accounted for the individual fairness loss function $L'_{ind}(h_2(f(X)))_D$.

It is worth mentioning that Equation 9 and 10 also give an insight of how to extend the demographic parity gap and yNN to multi-outcome and multi-group datasets.

5 EXPERIMENTS AND ANALYSIS

In this section, we conducted extensive experiments to evaluate the effectiveness of GIFair compared with baseline algorithms. All experiments were conducted on a machine with 3.6GHz CPU and 32GB memory. We implemented all algorithms in Python.

5.1 Datasets

We conducted experiments on 3 widely used real datasets, COMPAS, Adult and German. The statistics of datasets are listed in Table 1.

COMPAS collected by ProPublica [2] contains criminal offense information of 6k individuals. Each instance contains 12 attributes, including age, race, count of prior crimes, etc. Dataset COMPAS is often used to predict whether a criminal defendant will recidivate. For this dataset, we use *race* ($A = 1$ for African-Americans and $A = 0$ for other races) as the protected attribute.

Adult collected by UCI [7] contains 45k instances of information describing adults (e.g., gender and native country). Each instance in the dataset consists of 14 attributes. We use this dataset to predict each person's income category ($Y = 1$ if the income is greater than 50K per year, and $Y = 0$ otherwise). We use attribute *gender* ($A = 1$ for females and $A = 0$ for males) as the protected attribute.

German collected by UCI [7] contains the information of 1k individuals, each of which is described by 20 attributes (e.g., age and credit history). This dataset classifies each individual as good or bad credit risks ($Y = 0$ for good credit risks and $Y = 1$ for bad credit risks). We use attribute *age* ($A = 1$ for the aged and $A = 0$ for the young) as the protected attribute.

5.2 Baseline Algorithms

We selected the following algorithms as baseline methods. (1) UNFAIR [30]: a normal classification algorithm that does not consider fairness. (2) ALFR [9]: aims at learning flexible representations that are demographic parity. (3) LAFTR [22]: includes three variants, which target demographic parity, equalized odds, and equal opportunity. (4) DCFR [30]: uses three variants to solve demographic parity, equalized odds and conditional fairness. (5) LFR [32]: aims at generating a representation that achieves both group fairness, and individual fairness with a non-adversarial learning approach.

In this work, demographic parity is used as group fairness model. For comparison, we compared all baselines with demographic parity only, and thus in the experiments, we only use the variants of LAFTR and DCFR that consider demographic parity as baselines.

5.3 Measurements

To compare the performance of algorithms, we used the following three metrics widely adopted in the literature. (1) *Accuracy* (denoted by *ACC*): measuring the difference between the outcome y_i and the predicted outcome $\hat{y}_i$ of all data points x_i, i.e., $ACC = 1 - \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$. (2) *Demographic parity gap* (denoted by ΔDP): measuring the group fairness (introduced in Equation 1). (3) *yNN* (denoted by ΔyNN): measuring the individual fairness (introduced in Equation 2).

5.4 Parameter Setting

By default, GIFair in the experiments uses the original loss function (defined in Equation 6) which has three coefficients α, β and δ to control the weight of accuracy, group fairness and individual fairness, respectively, and thus, the original loss function is more suitable for studying the trade-off among these three targets, which is the major focus in this paper. Note that, by default, the focal loss function (defined in Equation 8) is not used because it only has one parameter γ to achieve an overall balance among the three targets, which does not closely explore the detailed trade-off among them.

Specifically, to study the trade-off between accuracy and group fairness in GIFair, we fixed both α and δ to 1 and changed β from

0.1 to 20. Besides, we changed δ from 0.1 to 20 to study the trade-off between accuracy and individual fairness where α and β are both fixed to 1. We also studied the trade-off between two kinds of fairness by varying β and δ from 0.1 to 20, respectively, and fixing other coefficients to 1. For baseline algorithms, we also changed their coefficients from 0.1 to 20. When testing the performance of the focal loss function, we varied the value of focusing parameter γ from 0.05 to 5. 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. Implementation details of algorithms can be found in Section A.

5.5 Results

5.5.1 Trade-off Studies. We studied the trade-off between any two terms from accuracy, group fairness and individual fairness. According to [30], adversarial representation learning has become the state-of-the-art method of the fair classification problem. For comparison, we selected the algorithms using the method of adversarial representation learning to study the trade-off of algorithms. Note that the non-adversarial learning baseline (i.e., LFR) performs worse than GIFair and other baselines for all the three measurements and all the datasets, the details shown in Section B.1. We compared the performance of algorithms by computing the Pareto front curves of different algorithms as shown in Figure 2. Since algorithm UNFAIR does not have weights for trading-off, we used a star mark for it. We also use a diamond mark for an "ideal" model in each figure, representing the preferable performance. Specifically, the "ideal" model is assumed to obtain the best performance for each measurement observed in the experiments.

Accuracy and Group Fairness. Figure 2(a) and (b) show the trade-off between accuracy and group fairness on datasets COMPAS and Adult, where the left-top points (high ACC, low ΔDP) are preferable.

On dataset COMPAS, we notice that in the range [0.005, 0.071] of ΔDP, GIFair can achieve a much better accuracy performance than other baseline algorithms under the same ΔDP performance. GIFair is said to dominate all other baseline algorithms in this range. Besides, the range of small ΔDP (which means good group fairness performance) is the most important range for fair classification algorithms. GIFair achieves the leftmost point ($\Delta DP = 0.0001$). It shows that GIFair can achieve the group fairest results, which is an important property for fair classification algorithms.

The comparison results are similar on dataset Adult, and GIFair still dominates baselines in range [0, 0.018] and [0.12, 0.16] of ΔDP. GIFair also obtains the smallest ΔDP, 0.0019. Thus, GIFair can achieve a better trade-off between accuracy and group fairness.

About the relation between accuracy and group fairness, only when the dataset contains discrimination, accuracy and fairness are opposing. For example, in dataset COMPAS, the positive rate of African-Americans is far more than the positive rate of other races. For high accuracy, the predicted positive rate of one race will be far larger than other races. However, accuracy will be lost when achieving good group fairness. In this case, we could balance the fairness and accuracy by adjusting their weights, e.g., if the user prefers a classifier with better group fairness, he/she should give the weight β a larger value compared α and δ. Users could also set the values of weights by the experiment results. Each data point in the Pareto front curves corresponds to one setting of weights, and thus users could select the setting based on the performance of each data point.

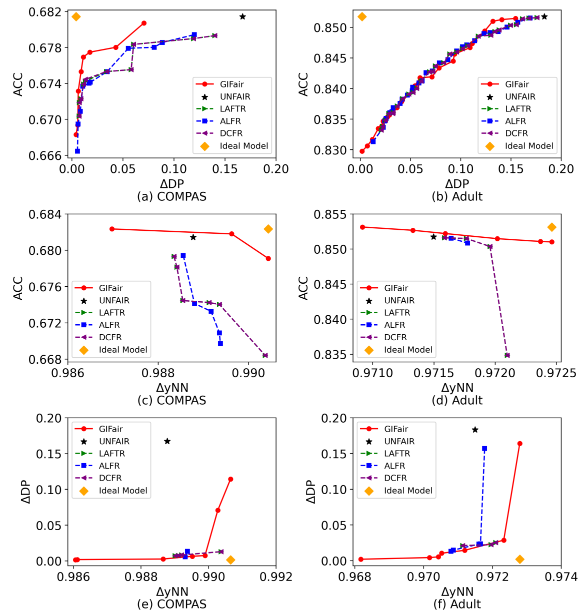

Figure 2: Trade-off Curves on Dataset COMPAS and Adult

Accuracy and Individual Fairness. Figure 2(c) and (d) show the trade-off between accuracy and individual fairness. The upper-right points (high ACC, high ΔyNN) are preferable. On both datasets COMPAS and Adult, GIFair performs much better than all state-of-art baselines, because GIFair can achieve better accuracy under the same performance of ΔyNN. For example, on dataset COMPAS, when the ΔyNN performance of all algorithms is 0.9887, the ACC performance of GIFair is 0.683, which is about 1% higher than other state-of-art baselines. Besides, we notice that the ACC performance of baselines will suddenly drop when ΔyNN increases since they do not optimize individual fairness, which means that when baseline algorithms want to pursue good individual fairness performance, their accuracy will be lost a lot. However, the ACC performance of GIFair is slightly affected by ΔyNN.

Group Fairness and Individual Fairness. We studied the trade-off between the two kinds of fairness in Figure 2(e) and (f). The bottom-right points (high ΔyNN, low ΔDP) are preferable. We observe that, on dataset COMPAS, the performance of different algorithms (except UNFAIR) is close. However, GIFair achieves the fairest result on both group and individual fairness. On dataset Adult, GIFair dominates all baselines and achieves the largest ΔyNN, 0.9728, and also the smallest ΔDP, 0.0018, which shows that GIFair can better trade group fairness off against individual fairness.

Moreover, GIFair obtains similar superiority on the trade-off among accuracy, group and individual fairness for dataset German compared with all the baselines. Due to lack of space, the details are presented in Section B.2.

5.5.2 Ablation Studies and Focal Loss Function. We conducted ablation studies for each of the two adversaries in GIFair and we also compared the original GIFair (using the original loss function) with a variant of GIFair (using the focal loss function). Specifically, we

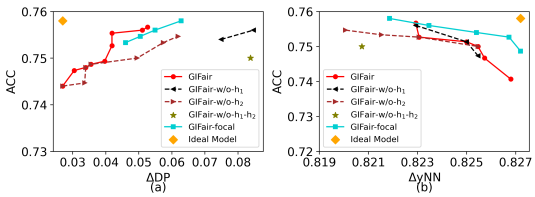

Figure 3: Ablation Studies and Focal Loss Function on Dataset German

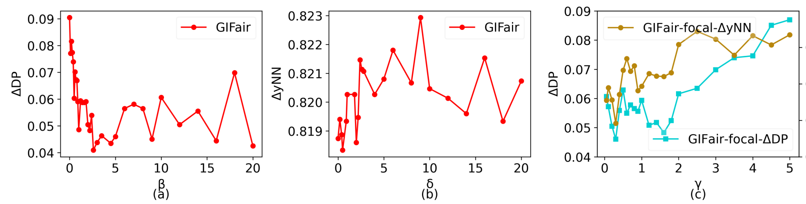

Figure 4: Parameter Studies on Dataset German

form the following variants. (1) GIFair without group adversary h_1 (denoted by **GIFair-w/o-h_1**), by setting coefficient β to 0 (and thus Step 1 of training h_1 in Algorithm 1 is skipped). (2) GIFair without individual adversary h_2 (denoted by **GIFair-w/o-h_2**), by setting coefficient δ to 0 (and thus Step 2 of training h_2 in Algorithm 1 is skipped). (3) GIFair without both h_1 and h_2 (denoted by **GIFair-w/o-h_1-h_2**), by setting both β and δ to 0 (and thus it becomes a normal classifier only optimizing the accuracy). (4) The variant of GIFair using the focal loss function (denoted by **GIFair-focal**).

Figure 3 illustrates the results of ablation studies on dataset German. Without group adversary h_1, GIFair-w/o-h_1 has much larger ΔDP (i.e., worse group fairness) than GIFair and other variants on the same high accuracy level (i.e., when ACC is around 0.755, ΔDP of GIFair-w/o-h_1 and GIFair are around 0.08 and 0.04 shown in Figure 3(a)). This verifies the effectiveness of improving group fairness using the group adversary. Similarly, for a high accuracy ($ACC \approx 0.755$), GIFair has larger ΔyNN (0.823) than GIFair-w/o-h_2 (0.82) shown in Figure 3(b), indicating that the individual adversary h_2 could effectively improve individual fairness. Besides, it is clear from Figure 3(a) and (b) that, without both adversaries, GIFair-w/o-h_1-h_2 obtains bad performance for both group and individual fairness.

When the focal loss function is used, GIFair-focal achieves an even better trade-off between accuracy and individual fairness (i.e., a larger ΔyNN when ACC in $[0.75, 0.755]$, shown in Figure 3(b)). This is because the two terms receive more balanced weights for GIFair-focal, but GIFair tries much large weights to achieve invidual fairness which slightly sacrifices accuracy. As shown in Figure 3(a), although GIFair-focal fails to achieve as good group accuracy as GIFair, the highest accuracy is obtained for GIFair-focal (i.e., $ACC = 0.758$). Thus, we conclude that using focal loss function is effective to improve individual fairness and accuracy for GIFair.

5.5.3 *Case Studies.*

We conducted case studies for the classification results regarding group fairness and individual fairness.

When only individual fairness is optimized (i.e., setting group fairness coefficient β to 0) for dataset COMPAS, we observe a representative prediction result where 47% of the African-American group will recidivate, while this proportion for the group containing 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 a much fairer result.For individual fairness, in dataset COMPAS, there exist some pairs of similar defendants who only have 1 day difference on attribute *days_b_screening_arrest* (i.e., the days between screening and arrest) and have the same value for all other attributes. When only group fairness is optimized (i.e., setting individual fairness coefficient 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. Similar case study results for the other datasets can be found in Section B.4.

5.5.4 *Parameter Studies.*

We studied the effect of β for group fairness, δ for individual fairness and γ for the focal loss function in GIFair (note that coefficient α for accuracy is fixed to 1).

As shown in Figure 4(a) and (b) for dataset German, when β (resp. δ) increases, the group (resp. individual) fairness first obtains better performance with lower ΔDP (resp. higher ΔyNN), because setting β (resp. δ) higher means group (resp. individual) fairness is more focused. However, the performances of both fairness measurements become unstable after β and δ are larger than 5, which indicates that an "over-focused" weight is given to each fairness target causing the model to be over-fitting. This suggests that β and δ should not be set to a too large value to obtain a steady performance.

When γ for the focal loss function is increased, higher ΔDP (i.e., worse group fairness) and higher ΔyNN (i.e., better individual fairness) are observed, shown in Figure 4(c). This is because, according to Equation 8, larger γ indicates less balanced parameter setting, and then the loss with smaller value (i.e., group fairness loss for dataset German in this case) receives an even smaller weight. This indicates that γ should also not be set to a too large value for a better balance among accuracy, group fairness and individual fairness.

We also study how the accuracy is affected by parameters β, δ and γ. When the three parameters are increased, the classification accuracy also drops since the weight for accuracy decreases correspondingly in all the three cases. The detailed results can be found in Section B.5.

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. We theoretically show the incompatibility of the two kinds of fairness and that GIFair guarantees the group fairness performance of any classifier trained on the representation. In our experiments on 3 real datasets, GIFair outperforms baselines with better fairness and higher accuracy. For future work, we would like to explore a more efficient individual fairness metric than yNN and to achieve a holistic optimization for utility and multiple fairness goals at the same time.

REFERENCES

[1] Agarwal, Sushant. 2020. Trade-Offs between Fairness, Interpretability, and Privacy in Machine Learning. *UWSpace* (2020).

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: Risk assessments in criminal sentencing. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Pranjal Awasthi, Corinna Cortes, Yishay Mansour, and Mehryar Mohri. 2020. Beyond individual and group fairness. *arXiv preprint arXiv:2008.09490* (2020).

[4] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. *JMLR* (2018).

[5] Sara N. Bleich, Mary G. Findling, Logan S. Casey, Robert J. Blendon, John M. Benson, Gillian K. SteelFisher, Justin M. Sayde, and Carolyn Miller. 2019. Discrimination in the United States: Experiences of black Americans. *HSR* 54, S2 (2019), 1399–1408. https://doi.org/10.1111/1475-6773.13220

[6] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *ICML*, Vol. 97. 1397–1405.

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

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *ITCS* (Cambridge, Massachusetts). 214–226. https://doi.org/10.1145/2090236.2090255

[9] Harrison Edwards and Amos Storkey. 2016. Censoring Representations with an Adversary. In *ICLR*.

[10] David García-Soriano and Francesco Bonchi. 2021. Maxmin-fair ranking: individual fairness under group-fairness constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 436–446.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*. 2672–2680.

[12] Dandan Guo, Chaojie Wang, Baoxiang Wang, and Hongyuan Zha. 2022. Learning Fair Representations via Distance Correlation Minimization. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[13] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS* (Barcelona, Spain). 3323–3331.

[14] F. Kamiran and T. Calders. 2009. Classifying without discriminating. In *ICCCC*. 1–6.

[15] F. Kamiran and T. Calders. 2011. Data preprocessing techniques for classification without discrimination. In *KAIS*, Vol. 33. 1–33.

[16] Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. 2022. Learning fair representation with a parametric integral probability metric. *arXiv preprint arXiv:2202.02943* (2022).

[17] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In *ICDE*. 1334–1345. https://doi.org/10.1109/ICDE.2019.00121

[18] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. Operationalizing Individual Fairness with Pairwise Fair Representations. In *VLDB*, Vol. 13. 506–518. https://doi.org/10.14778/3372716.3372723

[19] Xuran Li, Peng Wu, and Jing Su. 2022. Accurate Fairness: Improving Individual Fairness without Trading Accuracy. *arXiv preprint arXiv:2205.08704* (2022).

[20] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. *ICCV* (2017), 2999–3007.

[21] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. 2022. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1088–1097.

[22] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*, Vol. 80. 3384–3393.

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

[24] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. 2022. Learning Fair Representation via Distributional Contrastive Disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1295–1305.

[25] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. 2022. On the Impossibility of Non-trivial Accuracy in Presence of Fairness Constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7993–8000.

[26] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *SIGMOD*. 793–810. https://doi.org/10.1145/3299869.3319901

[27] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *NeurIPS*, Vol. 32. 5427–5437.

[28] Sahil Verma, Michael Ernst, and Rene Just. 2021. Removing biased data to improve fairness and accuracy. *arXiv preprint arXiv:2102.03054* (2021).

[29] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. 2021. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1748–1757.

[30] Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen, and Wei Cui. 2020. Algorithmic Decision Making with Conditional Fairness. In *KDD*. 2125–2135. https://doi.org/10.1145/3394486.3403263

[31] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *CIKM* (Singapore, Singapore). 1569–1578. https://doi.org/10.1145/3132847.3132938

[32] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *ICML*, Vol. 28. 325–333.

[33] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. 2020. Conditional Learning of Fair Representations. In *ICLR*.

A IMPLEMENTATION DETAILS

The two adversaries of GIFair are both feedforward neural networks with a single hidden layer, which has 8 units on dataset COMPAS, 50 units on dataset Adult and 4 units on dataset German. We trained our model for 500 epochs and then, fine-tuned it. We adopted Adadelta as the optimizer of which the learning rate is 1. The batch size is 256 for dataset COMPAS, 512 for dataset Adult and 64 for dataset German. The dimensionality of representation Z is 8 for dataset COMPAS, 60 for dataset Adult and 40 for dataset German.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 Overall Performance

Table 2 shows the most representative results of all algorithms (including the variant of GIFair that enables the focal loss function) for four choices of best performance: (a) the result with the best accuracy (ACC), (b) the result with best group fairness (ΔDP), (c) the result with best individual fairness (ΔyNN), (d) the result with the largest *sum* (where *sum* is defined to be $ACC + (1 - \Delta DP) + \Delta yNN$). We notice that GIFair outperforms baseline algorithms on the four choices in most cases. For example, on choice (a), GIFair achieves the largest value of ACC, 0.6818, on COMPAS among all algorithms. GIFair also achieves the smallest value of ΔDP, 0.0051 (on choice (b)) and the largest value of ΔyNN, 0.9931 (on choice (c)) on COMPAS among all algorithms. Similar superior performance of GIFair can be found on the other two datasets, Adult and German. Besides, we compared the three measurements at the same time on choice (d). For all the three datasets, GIFair achieves the best group fairness (i.e., smallest ΔDP). We also observe that GIFair-focal achieve best accuracy on datasets COMPAS and Adult, which shows that GIFair-focal is capable of achieving the accurate results while the three targets are optimized in a balanced way. The aboves results illustrate that our algorithm GIFair can achieve the best overall results.

B.2 Remaining Trade-off Studies

On dataset German, we also observe the dominating performance of GIFair on the trade-off between accuracy and group fairness (shown in Figure 5(a)) and on the trade-off between group fairness and individual fairness (shown in Figure 5(c)). Specifically, our GIFair algorithm achieves ΔDP (for group fairness) in $[0.01, 0.12]$ for accuracy in $[0.756, 0.767]$, while the best-performed baseline has ΔDP at least 0.04 and accuracy at most 0.762. Also, the ΔyNN (for individual fairness) reaches around 0.84, which outperforms all other baselines, as shown in Figure 5(b) and (c).

Table 2: Comparison of GIFair and Baseline Algorithms

Tuning	Method	COMPAS			Adult			German		
		ACC	ΔDP	ΔyNN	*ACC*	ΔDP	ΔyNN	*ACC*	ΔDP	ΔyNN
Baseline	UNFAIR	0.6798	0.1660	0.9890	0.8510	0.1858	0.9728	0.7500	0.0838	0.8307
(a) Max *ACC*	LFR	0.6510	0.1410	0.9499	0.8517	0.1056	0.9610	0.7201	0.0683	0.7845
	ALFR	0.6792	0.1197	0.9883	0.8514	0.1707	0.9722	0.7613	0.0543	0.8266
	LAFTR	0.6795	0.1403	0.9884	0.8514	0.1760	0.9719	0.7587	0.0670	0.8269
	DCFR	0.6795	0.1403	0.9884	0.8514	0.1760	0.9719	0.7587	0.0670	0.8269
	GIFair	**0.6818**	0.1552	0.9865	**0.8523**	0.1844	0.9714	**0.7667**	0.0111	0.8080
	GIFair-focal	0.6754	0.0197	0.9875	0.8467	0.0947	0.9714	0.7633	0.0667	0.8320
(b) Min ΔDP	LFR	0.5867	0.0424	0.9586	0.7186	0.0173	0.9641	0.7198	0.0574	0.7832
	ALFR	0.6665	0.0054	0.9879	0.8309	0.0144	0.9700	0.7540	0.0451	0.8213
	LAFTR	0.6703	0.0072	0.9892	0.8309	0.0168	0.9690	0.7527	0.0483	0.8270
	DCFR	0.6703	0.0072	0.9892	0.8309	0.0168	0.9690	0.7527	0.0483	0.8270
	GIFair	0.6707	**0.0051**	0.9880	0.8316	**0.0092**	0.9700	0.7567	**0.0000**	0.8073
	GIFair-focal	0.6711	0.0120	0.9876	0.8457	0.0917	0.9716	0.7500	0.0286	0.8197
(c) Max ΔyNN	LFR	0.5667	0.1004	0.9713	0.7186	0.0173	0.9641	0.7126	0.0679	0.8033
	ALFR	0.6728	0.0140	0.9893	0.8325	0.0188	**0.9732**	0.7547	0.0670	0.8324
	LAFTR	0.6725	0.0110	0.9901	0.8509	0.1477	0.9721	0.7520	0.0657	0.8329
	DCFR	0.6725	0.0110	0.9901	0.8509	0.1477	0.9721	0.7520	0.0657	0.8329
	GIFair	0.6700	0.0167	**0.9931**	0.8491	0.1239	0.9719	0.7433	0.0571	**0.8387**
	GIFair-focal	0.6704	0.0180	0.9881	0.8457	0.0917	0.9716	0.7633	0.0667	0.8320
(d) Max *sum*	LFR	0.5778	0.0429	0.9699	0.7121	0.0207	0.9629	0.7184	0.0661	0.8005
	ALFR	0.6709	0.0099	0.9891	0.8325	0.0188	**0.9732**	**0.7613**	0.0543	0.8266
	LAFTR	0.6703	0.0072	**0.9892**	0.8309	0.0168	0.9690	0.7533	0.0498	**0.8307**
	DCFR	0.6703	0.0072	**0.9892**	0.8309	0.0168	0.9690	0.7533	0.0498	**0.8307**
	GIFair	0.6707	**0.0051**	0.9880	0.8316	**0.0092**	0.9700	0.7600	**0.0016**	0.8117
	GIFair-focal	**0.6731**	0.0124	0.9880	**0.8457**	0.0917	0.9716	0.7600	0.0397	0.8213

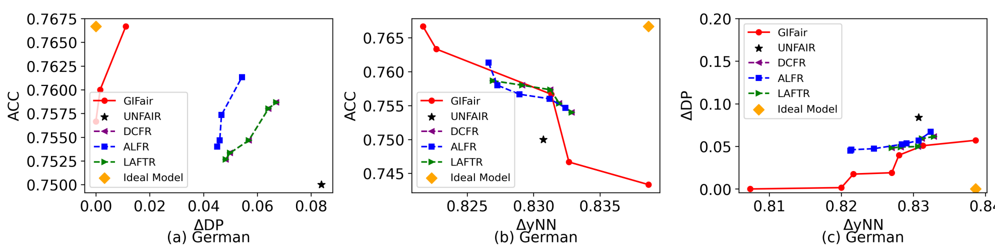

Figure 5: Trade-off Curves on Dataset German

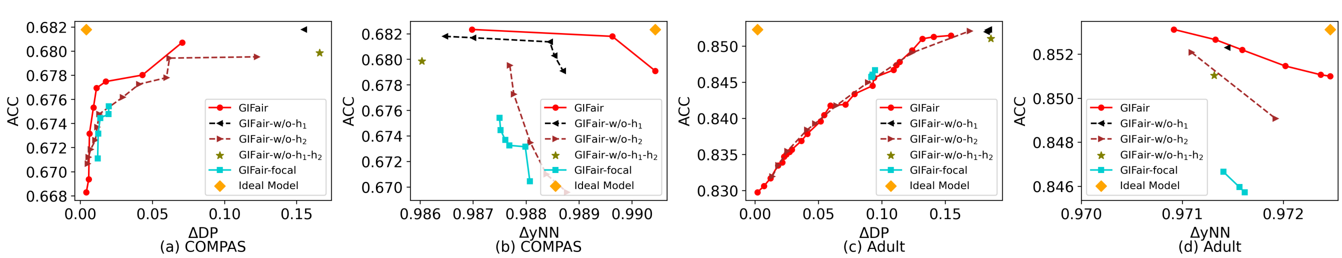

Figure 6: Ablation Studies and Focal Loss Function on Dataset COMPAS and Adult

B.3 Remaining Ablation Studies and Focal Loss Function

We show the remaining results for the ablation studies and the comparison between the switching of focal loss function and the

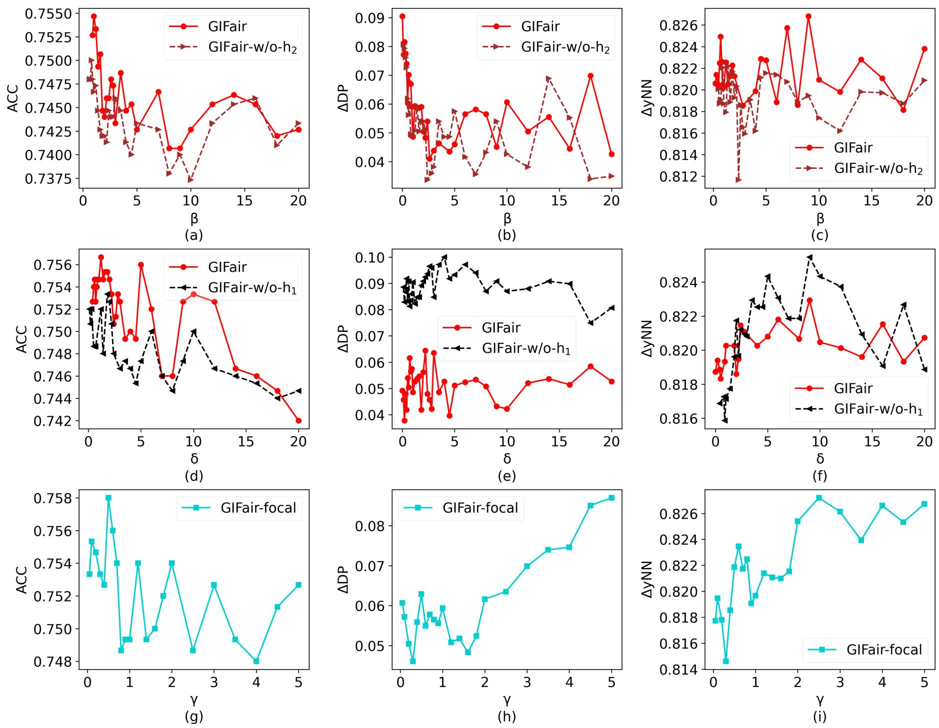

Figure 7: Effect of Parameters on Dataset German

original loss function on datasets COMPAS and Adult in Figure 6. As shown in Figure 6(a), when the group adversary is not included (i.e., GIFair-w/o-h_1), the ΔDP performance degrades significantly for dataset COMPAS. Similarly, when the individual adversary is not included (i.e., GIFair-w/o-h_2), the ΔyNN performance varies in a smaller range than GIFair shown in Figure 6(b) for dataset COMPAS. In Figure 6(c) and (d), we have similar conclusion on dataset Adult. Overall, the full GIFair obtain the superior fairness results when trading-off the accuracy and the two kinds of fairness due to the effectiveness of enabling the group adversary and individual fairness. Besides, we observe that only on dataset Adult, GIFair-focal obtains slightly better trade-off between accuracy and group fairness, as shown in Figure 6(c). It indicates that enabling the focal loss function does not always improve the trade-off results. Using the original loss function which searches through more combinations of coefficients could lead to better trade-off results in some cases.

B.4 Remaining Case Studies

We show similar case study results for dataset Adult and German.

Without optimizing group fairness for dataset Adult (i.e., setting group fairness coefficient β to 0), only 8.5% among the female group are predicted to have high income (i.e., > 50K per year), but this proportion is 26.7% among the male group. When both group and individual fairness are optimized, the high-income proportions among the female group and the male group are predicted to be 18.4% and 18.7%, respectively. Without optimizing individual fairness for dataset Adult (i.e., setting individual fairness coefficient β to 0), we found 16 pairs of similar adults who only have 2 hours difference on attribute *hours-per-week* (and have the same value for all other attributes) and are given different predictions. When both group and individual fairness are optimized, only 2 such pairs are found.

When only individual fairness is optimized (i.e., setting group fairness coefficient β to 0) for dataset German, one representative trained classifier predicts that 81.1% of the aged group may have bad credit risks, while 70% of the young group may have bad credit risks. When both group and individual fairness are optimized, it is improved to a fairer result where the bad credit risks proportions among the aged and young are predicted to be 75.6% and 74.3%, respectively. Regarding individual fairness, since dataset German has a relatively small data size, there do not exist any pair of closely similar individuals. However, according to our GIFair model using the Euclidean distance to measure the dissimilarity between two individuals, some similar pairs are found, e.g., two individuals who have small difference in 3 attributes only (i.e., *duration_in_month*, *credit_amount* and *present_residence_since*) and are the same for all other attributes. When only group fairness is optimized (i.e., setting individual fairness coefficient δ to 0), we found 6 such pairs of similar individuals that obtain different predictions in a representative result. This number improves to 2 when both group and individual fairness are optimized.

B.5 Remaining Parameter Studies

We present the full results of our parameter studies. For each of the three datasets, we study the effect of each parameter among β (for

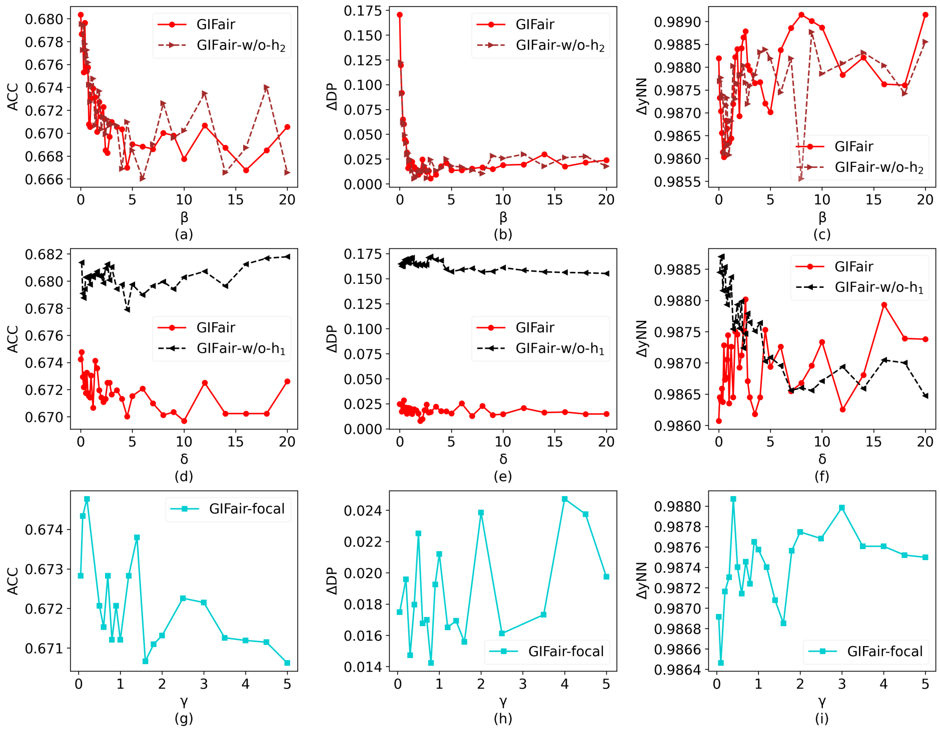

Figure 8: Effect of Parameters on Dataset COMPAS

group fairness), δ (for individual fairness) and γ (for the focal loss function) on each measurement among ACC, ΔDP and ΔyNN.

When β increases from 0.1 to 20 (other coefficients are fixed to 1 for GIFair), ACC drops first for all datasets since the accuracy receives less attention, and in turn the performance of group fairness improves significantly (ΔDP decreasing) as the weight of group fairness loss increases (see Figure 7(a), (b), Figure 8(a), (b) and Figure 9(a), (b)). For the two smaller-scaled dataset German and COMPAS, when β is set to a large value (e.g., > 5), the performance of accuracy and fairness are easier to be unstable. This is because the over-fitting due to over-large weight (as we introduced in Section 5.5.4) could be much obvious on small datasets. For the larger dataset Adult, the performance are less sensitive to the change of β. And thus, we can more easily observe the trend of decreasing ΔyNN (see Figure 9(c)), since the individual fairness also obtains less attention when β is increased. Besides, we also compare GIFair with GIFair-w/o-h_2 (by fixing δ to 0) in this case. The overall trends of the two algorithms are similar, but GIFair-w/o-h_2 has more unstable performance for larger β because the over-fitting effect is more obvious (since β has a more over-large weight in this case).

When δ increases from 0.1 to 20 (other coefficients are fixed to 1 for GIFair), we observe similar trends of decreasing the accuracy and increasing individual fairness (see Figure 7(d), (f), Figure 8(d), (f) and Figure 9(d), (f)). However, the change of δ does not obviously affect the performance of group fairness on dataset COMPAS (as shown in Figure 8(e)) and dataset Adult (as shown in Figure 9(e)).

When γ increases from 0.05 to 0.5, we observe decreasing accuracy, increase ΔDP (worse group fairness) and increasing ΔyNN (better individual fairness) for all three datasets (see Figure 7(g), (h), (i), Figure 8(g), (h), (i) and Figure 9(g), (h), (i)). This is because, for all the datasets, the range of ΔyNN values is the largest and range of ΔDP values is the smallest. Thus, as γ is set larger in the focal loss function, the balancing among the three targets is less retained, and then the accuracy and the group fairness will receive less attention than a smaller γ value, while the individual fairness will obtain more attention.

C PROOFS OF THEOREMS

Proof of Theorem 4.1. First, we show when either of the two conditions is satisfied by η, then η satisfies demographic parity and the yNN condition simultaneously. Condition 1: Since η is a constant function, the outcome of every data point is the same, i.e., $\eta(x_i) = \eta(x_j), \forall x_i, x_j \in D$, which means the yNN condition is satisfied. We can also get $Pr[\eta(x_i) = 1 | x_i \in D_1] = Pr[\eta(x_i) = 1 | x_i \in D_0] = 1$ or 0, so classifier η satisfies demographic parity. Condition 2: (1) The outcomes of data points in a k-NN cluster C are the same and the k-NN set of every data point in C also belongs to C, and thus the yNN condition is also satisfied. (2) Demographic parity is satisfied by definition.

Next, we show when η satisfies demographic parity and the yNN condition simultaneously, then either of the two conditions is satisfied by η. Consider a classifier η that does not satisfy Condition 1 or Condition 2. Condition 1 is violated which implies there exists two items $x_i, x_j \in D$ such that $\eta(x_i) \neq \eta(x_j)$. For Condition 2: Assume (1) is violated, i.e., there exists a k-NN cluster, in which the outcomes of two data points x_i, x_j are not the same. Since a

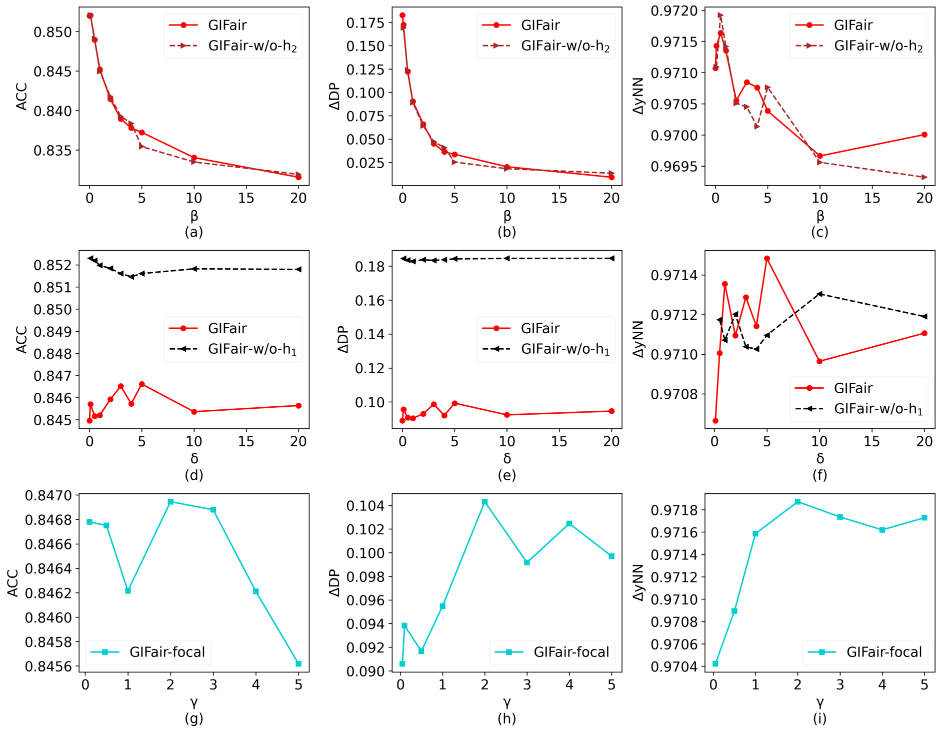

Figure 9: Effect of Parameters on Dataset Adult

k-NN cluster is a connected component, we can always find a path from x_i to x_j. Because the outcomes of x_i, x_j are different, we can find an edge in this path that connects two data points x'_i, x'_j and the outcomes of these two data points are different, which contradicts the definition of the yNN condition (i.e., a data point has the same outcome as all its nearest neighbors). Assume (2) is violated, demographic parity is not satisfied by definition. □

Proof of Theorem 4.2. Note that each $z_i \in Z$ has the same group membership as x_i. Then the demographic parity gap of η on Z, i.e., $\Delta DP_Z(\eta)$, is formalized as follows.

$$\Delta DP_Z(\eta) = \left| \frac{\sum_{x_i \in D_1} \eta(f(x_i))}{|D_1|} - \frac{\sum_{x_j \in D_0} \eta(f(x_j))}{|D_0|} \right| \tag{11}$$

It is easy to observe that $\Delta DP_Z(\eta)$ has the same form as $L_{grp}(h_1(Z), A)_D$, and thus we consider a group adversary h'_1 that always achieves the same result with η, i.e., $h'_1 = \eta$. Clearly, $L_{grp}(h'_1(Z), A)_D = \Delta DP_Z(\eta)$. Since the objective value of optimal group adversary h^*_1 is no less than the objective value of any h'_1, we can obtain $L_{grp}(h^*_1(Z), A)_D \geq L_{grp}(h'_1(Z), A)_D = \Delta DP_Z(\eta)$. □