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

Anonymous Author
Anonymous Department
Anonymous Organization
City, Country
email address or ORCID

Anonymous Author
Anonymous Department
Anonymous Organization
City, Country
email address or ORCID

Anonymous Author
Anonymous Department
Anonymous Organization
City, Country
email address or ORCID

***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, GIFair makes sure that the classifiers trained on the generated representation well reconcile both individual fairness and 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.**

***Index Terms*—Fairness, Adversarial Learning, Learning Representation**

I. Introduction

Machine learning models are increasingly utilized for decision-making nowadays. Most classification tasks prioritize achieving high utility, but the fairness of machine learning models is often overlooked by researchers. Many studies have shown that using historical datasets that are biased against some groups of people to train machine learning models for high accuracy can lead to discrimination in the decision-making process. For instance, due to racism in the US criminal justice system, African-Americans are arrested and convicted at higher rates than other races, resulting in biased predictions of recidivism if these criminal records are used [1].

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 manifest 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, 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 [2]. Similarly, as shown in a study in 2017 involving 3453 American adults, African-Americans have a significant portion reporting discrimination [3]. These examples show the existence of systemic discrimination and thus we need to address racism or sexism more actively. 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* [4], [5] and *individual fairness* [6], [7]. 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 should be *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 directly process the historical datasets to mitigate discrimination by modifying the original outcomes [8] or the attributes of data [9]. Some methods achieve fairness during training by imposing fairness as a regularizer [10], [11] or hard constraints [12]. Representation learning [13], [14] is another common approach, which transforms 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 but do not address the problem of reconciling group fairness and individual fairness *at the same time*. Individual fairness is also a very important aspect of fairness. Only satisfying group fairness may harm individual fairness, which could create discrimination. For example, according to [11], 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. 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 [11], [15]–[18] aim to achieve both individual and group fairness in their designs. Our most closely-related work is LFR [11]. LFR addresses the accuracy,

group fairness and individual fairness for classification by defining a loss function combining 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. However, the three terms in the objective function 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 [18] explores an alternative formation of individual fairness called *counterfactual fairness* [7] which requires treating individuals similarly to their 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 only ensures fair treatment for two *counterfactually* similar individuals, and thus it cannot guarantee general and stronger individual-level fairness for *any* two similar individuals. Some studies address a different task known as *fair ranking* [15], [17], which aims to rank the individuals without bias. In [17], an objective function of individual unfairness is minimized while some hard constraint for group fairness is satisfied for ranking. [15] reconciles utility and each of the two fairness notions separately. Both [17] and [15] do not study the reconciliation between individual and group fairness. Another work [16] attempts to achieve the compatibility of individual and group fairness by a data-driven model based on the received unfairness complaints. However, this model is limited by the difficulty of obtaining the data about unfairness complaints which hinders its practical use.

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 reconcile both group fairness and individual fairness in the learned representation. Due to the conflict of the two types of fairness [19], one has to trade group fairness off against individual fairness, and in this paper, we aim to obtain a better trading-off performance between them. 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. For group (fairness) adversary, we apply a more effective formation of target function (compared with the prior adversarial learning approaches [20]), which results in a more powerful group adversary that better guarantees group fairness in our structure. For individual (fairness) adversary, we form its target function with a metric called *yNN* based on k-nearest neighbors, which address the explicit individual fairness requirement of treating any similar individuals equally (compared with existing studies that either only address individual fairness implicitly [11] or explore the alternative counterfactual fairness form [18]). We propose a well-designed training algorithm to reconcile all concepts (i.e., accuracy, individual fairness and group fairness) in our structure. Compared to the existing adversarial learning studies that only 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.

To further optimize our GIFair approach, we apply focal loss to address the two imbalance issues. Firstly, we propose a focal loss function so that the three targets receive more balanced attention in our training algorithm. GIFair with focal loss function is verified to have better trade-off performance (e.g., 30% improvement of group fairness under the same level of individual fairness) compared with the original version of GIFair. Secondly, we propose a new individual fairness notion called *balanced yNN* (generalizing the existing notion yNN), where the nearest neighbors that are farther away will be given a smaller weight. We show that the new notion achieves a more balanced measurement of individual fairness.

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 summarized as follows.

- We design a novel structure of adversarial representation learning with two adversaries for group fairness and individual fairness, respectively, each with an effective target function.
- We design a training algorithm that can well reconcile the two adversaries in our structure. Ablation analysis is conducted to show its superiority.
- We propose a focal loss function to ensure balanced attention of two types of fairness and accuracy.
- We propose a new individual fairness notion that measures individual fairness in a more balanced way.
- 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 II reviews related work. Section III presents the preliminaries. Section IV describes our solution to the fair classification problem. Section V presents the optimizations with focal loss. Then, Section VI reports experimental results and our analysis. Finally, Section VII concludes this paper.

II. Related Work

Most machine learning studies about fairness can be classified into three categories, namely *pre-processing*, *in-processing* and *post-processing*. *Pre-processing approaches* directly modify datasets to remove discrimination before using normal methods to do classification [9], [21]. *In-processing approaches*modify the classifier to improve its fairness performance while maintaining the utility of classification during training [10]–[12], [15]. 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, [10] probabilistically maps all items of the dataset into a low-rank representation that reconciles individual fairness and the utility of classifiers. Methods [12], [15] enforce fairness during training by modeling fairness as hard constraints. *Post-processing approaches*, e.g., [4], directly change the predicted outcomes of the learned predictors.

Learning Fair Representations. Recently, *fair representation learning* attracts great attention in fair machine learning, with LFR [11] (introduced in Section I) 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. [22] 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. [23] proposes distance covariance between the representation and the protected attribute as a new dependence measurement. DualFair [18] applies a contrastive self-supervised learning approach to obtain the representation satisfying both group fairness and counterfactual fairness. iFair [10] 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).

Among those approaches, adversarial representation learning has been broadly explored. ALFR [13] provides a framework to mitigate discrimination by learning representations that minimize the performance of the adversary that predicts the protected attribute of the representation. LAFTR [20] 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. IPM [24] 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 one type of fairness in all of the above studies of learning fair representations (except LFR [11] and DualFair [18]), we consider how to reconcile both group and individual fairness. As we introduced in Section I, both LFR and DualFair fail to address individual fairness effectively (since they do not adhere to the criterion of treating similar individuals similarly).

Trade-off between Accuracy and Fairness. Several previous studies have also explored the trade-off between accuracy and fairness. [25] improves the trade-off between group fairness and accuracy, on the problem of multi-task setting. [26] theoretically analyzes the difficulty of obtaining group fairness and accuracy simultaneously. [27] adapts a Siamese network approach to achieve the trade-off between accuracy and individual fairness. [15] targets the trade-off between accuracy and each of the individual and group fairness separately, but explores the problem of ranking. However, none of the above studies specifically addresses the trade-off between group fairness and individual fairness, which is the focus of our work.

III. Preliminaries

A. 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 an outcome attribute Y for classification and a protected attribute A representing the group membership (e.g., gender). Dataset D can thus be divided into different groups (e.g., females and males).

In line with most previous studies on fair classification [11], [13], [20], we assume binary outcome attribute and binary protected attribute (i.e., $Y \in \{0, 1\}$ and $A \in \{0, 1\}$). In our supplementary material [28], we discuss how we handle the multi-outcome and multi-group cases. 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 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.

B. Fairness Notions

Many fairness notions were proposed in the recent literature. For group fairness, two popular notions are *demographic parity* [5] and *equalized odds* [4]. 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 rates (FPR) and true positive rate (TPR) should also be equal among different groups. However, the above fairness notions may not be exactly satisfied by classifiers in most cases. Thus, following common approaches, we use *demographic parity gap* and *equalized odd distance* to measure how well a classifier satisfies group fairness. 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,

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

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 TPR of D_0 and the TPR of D_1, and the absolute difference between the 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 [6]. 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* [11], 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

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

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.

C. Generative Adversarial Network

Generative adversarial network (GAN) is an adversarial network [29] with two components, namely a *generator* G and a *discriminator* C. The generator G aims at deceiving the discriminator C by constructing synthetic data $G(z)$ 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 $G(z)$. Both components improve their ability through learning. That is, G is trained to generate $G(z)$ that cannot be distinguished from the real data, 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 to be $\mathbb{E}_{x \sim P_{data}}[\log(C(x))] + \mathbb{E}_{z \sim P_z}[1 - \log(C(G(z))]$. Specifically, discriminator C seeks to maximize $V(G, C)$ but generator G seeks to minimize $V(G, C)$.

IV. Methodology

A. 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.

Due to the conflict of group and individual fairness [19], the two fairness goals could not be satisfied simultaneously in most cases (an extended analysis on their incompatibility is given in our supplementary material [28]), and we need to balance the trade-off between them. We thus set our optimization goal of classifier η such that a balance can be obtained among accuracy, group fairness and individual fairness.

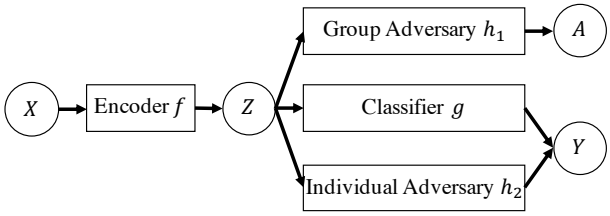

Fig. 1: Structure of GIFair

B. Model

First proposed by [13], plenty of existing works 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. 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 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 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 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 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 re-constructs 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: $\mathbb{R}^{d'} \to \{0, 1\}$ to predict the outcome $g(z_i)$ of each representation z_i in Z. Conceptually, let $g(Z) = g(f(X))$ denote the prediction process of all representations in Z. To preserve utility, we minimize a suitable classification loss function between $g(f(X))$ and Y in

a dataset D, denoted by $L_c(g(f(X)), Y)_D$ (which is selected to be cross-entropy in this work).

Group Adversary. To achieve group fairness of the generated representation Z, the group adversary h_1: $\mathbb{R}^{d'} \rightarrow \{0, 1\}$ 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 *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 $L_g(h_1(f(X)), A)_D$, as follows.

$$\begin{aligned} L_g(h_1(f(X)), A)_D &= |F_{D_0 \rightarrow 1}(h_1) - F_{D_1 \rightarrow 1}(h_1)| \\ &= |\frac{\sum\limits_{x_i \in D_0} h_1(f(x_i))}{|D_0|} - \frac{\sum\limits_{x_j \in D_1} h_1(f(x_j))}{|D_1|}| \end{aligned} \tag{3}$$

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

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

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* $L_g(h_1(f(X)), A)_D$ in encoder f.

Note that our GIFair model is general to other fairness notions. For instance, we could replace our group loss function with the loss function for equalized odds (following [20]) if equalized odds is the group fairness target.

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: $\mathbb{R}^{d'} \rightarrow \{0, 1\}$ is included. Specifically, for each representation $z_i = f(x_i) \in D$, h_2 predicts an outcome $h_2(z_i) \in \{0, 1\}$ 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 $L_i(h_2(f(X)))_D$, 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 $z_i \in Z$.

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

When $L_i(h_2(f(X)))_D$ 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* $L_i(h_2(f(X)))_D$ so 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 $L_i(h_2(f(X)))_D$ is *minimized*.

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. Note that we do not use 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. Using $f(X)$ for distance computation could be ineffective when encoder f distorts the similarity relationships among the data points.

Total Loss. The total loss function $L(f, g, h_1, h_2)_D$ 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.

$$\begin{aligned} L(f, g, h_1, h_2)_D = \alpha \cdot L_c(g(f(X)), Y)_D \\ + \beta \cdot L_g(h_1(f(X)), A)_D \\ + \delta \cdot L_i(h_2(f(X)))_D \end{aligned} \tag{5}$$

The coefficients α, β and δ provide a trade-off among accuracy, group fairness and individual fairness. Following the adversarial learning scheme, we train our model with a min-max optimization of our total loss function: $\min_{f,g} \max_{h_1,h_2} \mathbb{E}_{X,A,Y}[L(f, g, h_1, h_2)_D]$, where h_1 and h_2 are trained separately to maximize the total loss, while f and g are trained jointly to minimize the total loss.

Training. Algorithm 1 shows the pseudo-code of our learning algorithm of GIFair. Let θ_f, θ_g, θ_{h_1} and θ_{h_2} be the trainable parameters of f, g, h_1 and h_2, respectively. Our algorithm runs in e (an input integer) epochs. In each epoch, we first sample a mini-batch D' of size m (an input integer) from the dataset D. 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

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

1: **for** epoch from 1 to e **do**
2: Randomly sample a mini-batch D' from D of size m
3: ▷ *Step 1 (Train h_1)*
4: Freeze h_2, f, g; Unfreeze h_1
5: Optimize h_1 by ascending along gradient on D': $\nabla_{\theta_{h_1}} L_g(h_1(f(X)), A)_{D'}$
6: ▷ *Step 2 (Train h_2)*
7: Freeze h_1, f, g; Unfreeze h_2
8: Optimize h_2 by ascending along gradient on D': $\nabla_{\theta_{h_2}} L_i(h_2(f(X)))_{D'}$
9: ▷ *Step 3 (Train f and g)*
10: Freeze h_1, h_2; Unfreeze f, g
11: Optimize f and g by descending along gradient on D': $\nabla_{\theta_f, \theta_g} L(f, g, h_1, h_2)_{D'}$
12: **return** f, g, h_1, h_2

group adversary h_1 and individual adversary h_2, respectively, such that their objective functions are maximized. 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. 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.

C. Theoretical Properties of Loss Functions

We give the theoretical properties of our group loss function $L_g(h_1(f(X)), A)_D$ (Equation 3) and individual loss function $L_i(h_2(f(X)))_D$ (Equation 4), which show the effectiveness of using our loss functions to ensure fairness.

We first show that the optimal value of $L_g(h_1(Z), A)_D$ can upper-bound the demographic parity gap of any classifier trained on representation Z in Lemma IV.1 (note $Z = f(X)$). In the supplementary material [28], we provide the proofs.

Lemma IV.1. *For a group adversary h_1, the optimal value of $L_g(h_1(Z), A)_D$ (denoted by $L_g(h_1^*(Z), A)_D$) 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 IV.1, we connect $L_g(h_1(Z), A)_D$ 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 trained on Z given the optimal group adversary h_1^*. This shows the effectiveness of using $L_g(h_1(f(X)), A)_D$ as the group loss function.

Note that Lemma IV.1 differs from the theorem of bounding $\Delta DP_Z(\eta)$ in [20] due to different formation of group loss function. Our formation in Equation 3 captures both cases of h_1 predicting the group labels correctly and wrongly, where both cases lead to the effective adversary target as discussed in Section IV-B. The formation in [20] only covers the case where h_1 predicts the group labels correctly, and thus our formation is more effective and can still bound $\Delta DP_Z(\eta)$.

Analogously, we want to show the effectiveness of the individual loss function $L_i(h_2(Z))_D$. 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 nearest neighbors 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 IV.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 $L_i(h_2(Z))_D$.

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

In Lemma IV.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 function $L_i(h_2^*(Z))_D$ is effective.

V. Optimization with Focal Loss

To this end, we have formed our GIFair structure. However, we notice two issues due to imbalanced values. Firstly, the ranges of the three losses in Equation 5 have large differences (e.g., the value of $L_i(h_2(f(X)))_D$ 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. Secondly, since the yNN metric considers whether a data point x_i has the same prediction with its k nearest neighbors, it is possible that some of the nearest neighbors has large distances to x_i. Given two nearest neighbors x_j and $x_{j'}$ of x_i (where x_j has much smaller distance to x_i than $x_{j'}$), intuitively, the pair x_i and x_j should be given more attention than the pair x_i and $x_{j'}$ for contributing to the yNN metric.

To solve both issues, we exploit the idea of focal loss function [30] (originally for the class imbalance problem) to alleviate the imbalance among the three losses in our loss function and the imbalance of different distances for nearest neighbors. Consider an item with two possible outcomes, namely 1 and 0. Let p be the estimated probability that this item has 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 *Focal Loss 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 (i.e., close to 1), its weight $(1-p_t)^\gamma$ will be low (i.e., close to 0). In this way, the focal loss function could give less (resp. more) weights assigned to items with higher (resp. lower) p_t values.

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

$$\begin{aligned} & L(f,g,h_1,h_2)_D \\ &= (1-L_c(g(f(X)),Y)_D)^\gamma \cdot L_c(g(f(X)),Y)_D \\ &+ (1-L_g(h_1(f(X)),A)_D)^\gamma \cdot L_g(h_1(f(X)),A)_D \\ &+ (1-L_i(h_2(f(X)))_D)^\gamma \cdot L_i(h_2(f(X)))_D \end{aligned} \tag{6}$$

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.

Next, we show how we address the imbalance of varied distances for nearest neighbors. Similarly, our idea is to give a smaller (resp. larger) weight to a pair of data points whose pairwise distance is larger (resp. smaller) for computing the yNN metric by applying the *weight* term in the focal loss function. Specifically, since the focal loss function works for a variable between 0 and 1, we first define the *relative distance* between two data points x_i and x_j in the dataset D, denoted by $rd(x_i, x_j)$, to be the ratio of the Euclidean distance between x_i and x_j to the maximum Euclidean distance among all pairs of data points in D. Now, we propose a new metric called *Balanced yNN* to measure the individual fairness. Given a classifier η, a positive integer k, a non-negative number λ and dataset D, the balanced yNN of η for D, k and λ, denoted by $\Delta B\text{-}yNN_{D,k,\lambda}(\eta)$, is defined to be

$$\begin{aligned} & \Delta B\text{-}yNN_{D,k,\lambda}(\eta) \\ =& 1 - \frac{\sum_{x_i \in D} r_i \sum_{x_j \in \mathcal{N}_D^k(x_i)} (1-rd(x_i,x_j))^\lambda |\eta(x_i) - \eta(x_j)|}{k \cdot N} \end{aligned} \tag{7}$$

where $r_i = 1/\sum_{x_j \in \mathcal{N}_D^k(x_i)} (1-rd(x_i,x_j))^\lambda$ for each $x_i \in D$ is a re-weighting variable such that the maximum value of $\Delta B\text{-}yNN_{D,k,\lambda}(\eta)$ is always equal to 1. Clearly, the weight term $(1-rd(x_i,x_j))^\lambda$ has a smaller (resp. larger) value when $rd(x_i, x_j)$ is larger (resp. smaller).

Correspondingly, given the parameter λ, we also re-design our individual loss function $L_i(h_2(f(X)))_D$ as follows such that a better balanced yNN is obtained by GIFair.

$$\begin{aligned} & L_i(h_2(f(X)))_D \\ =& \frac{\sum_{x_i \in D} r_i \sum_{x_j \in \mathcal{N}_D^k(x_i)} (1-rd(x_i,x_j))^\lambda |h_2(f(x_i)) - h_2(f(x_j))|}{k \cdot N} \end{aligned} \tag{8}$$

Note that when λ is set to 0, the value of the weight term is always 1, and in this case the balanced yNN metric is equivalent to the original yNN. Therefore, the balanced yNN can be regarded as a generalization of yNN.

VI. Experiments and Analysis

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

TABLE I: 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

A. Datasets

We conducted experiments on 3 widely used real datasets, COMPAS, Adult and German. Table I lists the statistics.

COMPAS [1] 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 [31] contains 45k instances of information describing adults (e.g., gender and native country). Each instance 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 [32] 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.

B. Algorithms and Baselines

We selected the following algorithms as baseline methods. (1) UNFAIR: a normal classification algorithm that does not consider fairness. (2) LAFTR [20]: includes three variants, which target demographic parity, equalized odds, and equal opportunity. (3) LFR [11]: aims at generating a representation that achieves both group fairness, and individual fairness with a non-adversarial learning approach. (4) iFair [10]: aims at learning representations considering both accuracy and individual fairness. (5) DualFair [18]: learns representations satisfying both group fairness and counterfactual fairness.

If the original loss function defined in Equation 5 is used, our algorithm is denoted as GIFair, while GIFair-focal denotes our algorithm on the focal loss function defined in Equation 6.

C. Measurement

We focus on the classification accuracy, group fairness and individual fairness. (1) For accuracy, we use two widely adopted metrics, namely, *accuracy* (denoting ACC) defined to be the difference between the outcome y_i and the predicted outcome $\hat{y}_i$ of all data points x_i, i.e., $ACC = 1 - \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$, and the *F-1 score* (denoting $F1$) 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 III-B, namely *demographic parity gap*, denoted by ΔDP, and *equalized odds distance*, denoted by ΔEO. (3) For individual fairness, we use *yNN*, denoted by ΔyNN (introduced in Equation 2) and also our proposed *balanced yNN* metric, denoted by $\Delta B\text{-}yNN$ (introduced in Equation 7).

TABLE II: Comparison of GIFair and Baseline Algorithms

Dataset	Algorithm	Optimize Each Metric						Optimize Sum of Metrics					
		ACC	$F1$	ΔDP (↓)	ΔEO (↓)	ΔyNN	ΔB-yNN	ACC	$F1$	ΔDP (↓)	ΔEO (↓)	ΔyNN	ΔB-yNN
COMPAS	UNFAIR	0.6817	0.6328	0.1856	0.1504	0.9872	0.9882	0.6717	0.6228	0.1856	0.1504	0.9866	0.9882
	LFR	0.6510	0.5782	0.0424	0.0562	0.9713	0.9746	0.6510	0.5782	0.0424	0.0562	0.9713	0.9746
	LAFTR	0.6802	0.6325	0.0062	0.0244	0.9879	**0.9894**	0.6706	0.6186	0.0071	0.0244	**0.9872**	**0.9887**
	iFair	0.6801	0.6319	0.0493	0.0661	0.9878	0.9893	0.6780	0.6112	0.0512	0.0725	**0.9872**	0.9879
	DualFair	0.6810	0.6282	0.0124	0.0262	0.9793	0.9816	**0.6810**	0.6182	0.0124	0.0262	0.9793	0.9816
	GIFair	**0.6829**	**0.6344**	**0.0045**	**0.0231**	**0.9881**	**0.9894**	0.6699	0.6190	**0.0051**	**0.0231**	0.9870	0.9885
	GIFair-focal	0.6755	0.6300	0.0119	0.0344	0.9855	0.9871	0.6755	**0.6294**	0.0128	0.0344	0.9849	0.9865
Adult	UNFAIR	0.8510	0.8362	0.1858	0.1746	0.9728	0.9747	0.8310	**0.8162**	0.1858	0.1746	0.9728	0.9737
	LFR	0.8517	0.8332	0.0173	0.0374	0.9641	0.9684	0.7121	0.6932	0.0207	0.0382	0.9629	0.9697
	LAFTR	0.8514	0.8345	0.0168	0.0432	0.9721	0.9733	0.8309	0.8099	0.0168	0.0302	0.9690	0.9736
	iFair	0.8511	0.8339	0.0874	0.1263	**0.9730**	**0.9758**	0.8299	0.8068	0.0835	0.1293	**0.9729**	**0.9743**
	DualFair	0.8519	0.8334	0.0164	0.0306	0.9709	0.9720	0.8219	0.8134	0.0168	0.0335	0.9692	0.9711
	GIFair	**0.8523**	**0.8453**	**0.0092**	**0.0121**	0.9719	0.9736	0.8316	0.8154	**0.0092**	**0.0163**	0.9700	0.9715
	GIFair-focal	0.8467	0.8347	0.0917	0.1027	0.9716	0.9733	**0.8457**	0.8136	0.0917	0.1127	0.9716	0.9729
German	UNFAIR	0.7520	0.8273	0.0263	0.0253	0.7770	0.7932	0.7520	0.8273	0.0263	0.0253	0.7770	0.7932
	LFR	0.7201	0.7849	0.0683	0.0573	0.7845	0.8011	0.7201	0.7849	0.0683	0.0573	**0.7845**	0.8011
	LAFTR	0.7600	0.8320	0.0168	**0.0189**	0.7758	0.7918	0.7520	0.8264	0.0175	0.0189	0.7758	0.7918
	iFair	0.7596	0.8313	0.0404	0.0634	0.7898	0.8011	0.7514	0.8263	0.0425	0.0615	0.7819	**0.8014**
	DualFair	0.7601	0.8319	0.0183	0.0273	0.7845	0.7921	0.7501	0.8249	0.0147	**0.0173**	0.7745	0.7902
	GIFair	0.7620	0.8355	**0.0048**	0.0246	0.7863	0.8041	0.7547	**0.8297**	**0.0048**	0.0286	0.7813	0.7956
	GIFair-focal	**0.7667**	**0.8391**	0.0124	0.0253	**0.7935**	**0.8093**	**0.7587**	0.8332	0.0124	0.0266	0.7789	0.8010

D. Parameter Setting

We varied our parameters β and δ in our original loss function from 0.1 to 20, while parameter α is fixed to 1, to study the trade-off among the three targets. For baseline algorithms, we also changed their coefficients from 0.1 to 20. For the focal loss function, we varied the value of focusing parameter γ from 0.05 to 5. We also varied k from 2 to 10 when computing the k-nearest neighbors for yNN and λ from 0 to 20 for the balanced yNN. By default, we set k to 10 according to [10] and λ to 10 which is sufficiently large for the balance of pairwise distances (as verified later). 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 [28].

E. Results

1) Overall Comparison: Table II shows the overall comparison of our GIFair algorithm with all baselines. The left part in the table shows the best result for each metric when this metric is optimized alone. The right part shows the result when the sum of metrics is optimized. The sum of metrics is defined to be the sum of the accuracy metric (default to ACC), 1 minus the group fairness metric (default to ΔDP) and the fairness individual (default to ΔB-yNN). Note that for the group fairness metrics (i.e., ΔDP and ΔEO marked with "↓"), smaller values are preferred (thus we use 1 minus the group fairness metric for computing the sum of metrics), while all other metrics are favored with larger values.

As shown in Table II, GIFair and GIFair-focal outperform baseline algorithms in most cases. For example, GIFair reaches the best value for all metrics (when optimizing each metric) on dataset COMPAS. GIFair also obtains the best ΔDP and the best ΔEO for both dataset COMPAS and Adult. When the sum of metrics is optimized, GIFair-focal has the best $F1$ accuracy on dataset COMPAS and German and the best ACC on dataset Adult and German. On dataset German, GIFair-focal obtains the best individual fairness (of both ΔyNN and ΔB-yNN) when the two metrics are optimized.

2) 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 between different terms. We plotted the Pareto front curves (widely used in existing trade-off studies [10], [20]) for comparison. 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 each group fairness metric, so that for each figure, the right-top points (high values along each axis) are preferable. We show the results on one dataset (i.e., German), while we obtain similar results for the other two datasets, which are reported in our supplementary material [28].

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, although most algorithms could have a high accuracy of $ACC \approx 0.76$, 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, which shows much better group fairness. It is also observed that GIFair-focal could reach the highest ACC of around 0.765 but at a cost of sacrificing group fairness.

Accuracy and Individual Fairness. Figure 2(b) and (c) show the trade-off between accuracy and individual fairness, with ΔyNN and ΔB-yNN as the individual fairness metrics. GIFair and GIFair-focal still obtain the best trade-off, where GIFair-focal nearly converges to the optimal point. Similar trends are observed from these two figures. Specifically, when ACC is fixed to around 0.76, the baseline with the best indi-

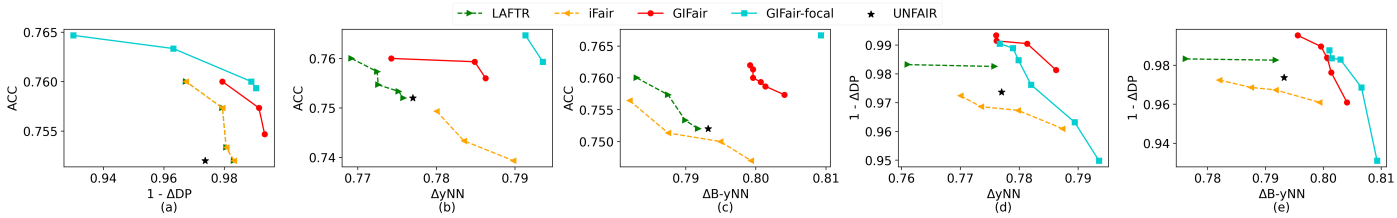

Fig. 2: Trade-off Curves on Dataset German

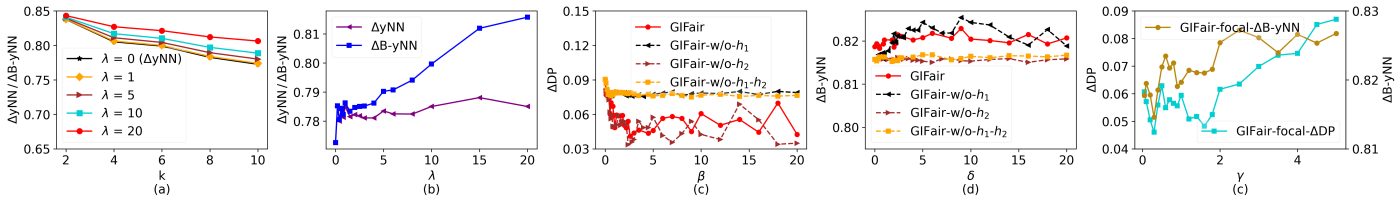

Fig. 3: Study of Parameters and Ablation Studies of GIFair on Dataset German

vidual 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 with its ACC below 0.74, while our GIFair-focal keeps ACC above 0.76, which improves iFair by more than 3%. In Figure 2(c), the performance of GIFair-focal dominate all baselines in both accuracy and individual fairness.

Group Fairness and Individual Fairness. Our algorithms also obtains superior trade-off between the two kinds of fairness, as shown in Figure 2(d) and (e), where we also explore two different individual fairness metrics. GIFair-focal achieves the highest ΔyNN (0.794) and the highest $\Delta B\text{-}yNN$ (0.809) among all algorithms, 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., $\Delta B\text{-}yNN = 0.804$), while its group fairness is only slightly downgraded (with $\Delta DP = 0.04$).

Insights from Trade-off Studies. Since we include multiple targets (i.e., accuracy and two types of fairness), one may not know which setting is the best suitable. The insights from our trade-off studies could address this issue. For instance, if a user wants a classifier of which the accuracy is at least 0.755 and group fairness performance is good, he/she could select the data point of which $ACC \geq 0.678$ and ΔDP is the smallest one in the curve that studies the trade-off between accuracy and group fairness. Then he/she could get the setting of weights (or the setting of focal loss function parameter γ) of this data point that corresponds to. Since our GIFair and GIFair-focal obtains the best trade-off among the three targets, the selected setting could give the user superior performance.

3) Ablation Studies: We conducted ablation studies for each of the two adversaries in GIFair. Specifically, we form the following variants. (1) GIFair without group adversary h_1 (denoted by **GIFair-w/o-**h_1), by skipping Step 1 of training h_1 in Algorithm 1. (2) GIFair without individual adversary h_2 (denoted by **GIFair-w/o-**h_2), by skipping Step 2 of training h_2 in Algorithm 1. (3) GIFair without both h_1 and h_2 (denoted by **GIFair-w/o-**h_1**-**h_2), by skipping both Step 1 and Step 2.

Figure 3 (c) and (d) illustrate 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 the original GIFair. This verifies the effectiveness of improving group fairness using the group adversary. Since group fairness is degraded, GIFair-w/o-h_1 has slightly better individual fairness than GIFair, because group fairness and individual fairness are two conflicting targets. Similarly, GIFair has larger $\Delta B\text{-}yNN$ than GIFair-w/o-h_2, indicating that the individual adversary h_2 could effectively improve individual fairness. Besides, without both adversaries, GIFair-w/o-h_1-h_2 obtains bad performance for both group and individual fairness.

4) Parameter Studies: We studied the effect of k for computing the k-nearest neighbors for yNN, λ for the balanced yNN, β for group fairness, δ for individual fairness and γ for the focal loss function in GIFair.

As shown in Figure 3(a), when k increases, each $\Delta B\text{-}yNN$ (for each λ) or ΔyNN (for $\lambda = 0$) decreases. This is because when considering more nearest neighbors, it is more likely that some nearest neighbors have a different prediction label to degrade individual fairness, which could be regarded as an "unstable" situation. Moreover, if we set a larger λ, the decrease of $\Delta B\text{-}yNN$ is less obvious, which reduces this "unstableness". The rationale is that using the balanced yNN can lower the influence of the far-away nearest neighbors (by assign them small weights). Therefore, we prefer a larger λ value.

We then fix k to 10. When trained with increasing λ, as shown in Figure 3(b), the resulting $\Delta B\text{-}yNN$ increases which accords with the trend in Figure 3(a). Interestingly, we also observe a slight increase of ΔyNN when λ increases from 0 to 10, which verifies the benefit of using the balanced loss function (introduced in Equation 8) for a suitably larger λ. Thus, we fix λ to 10 in all other experiments.

As shown in Figure 3(c) and (d), when β (resp. δ) increases, the group (resp. individual) fairness obtains better performance with lower ΔDP (resp. higher ΔyNN), because setting β (resp. δ) higher means group (resp. individual) fairness is more focused. 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 3(e). This is because, according to Equation 6, larger γ indicates less balanced parameter setting, and then the loss with larger value (i.e., group fairness loss for dataset German in this case) receives an even smaller weight. Similar results on the other two datasets can be found in our supplementary material [28].

5) 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 our supplementary material [28].

VII. 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.

References

[1] 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

[2] 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.

[3] S. Bleich, M. Findling, L. Casey, R. Blendon, J. Benson, G. SteelFisher, J. Sayde, and C. Miller, "Discrimination in the united states: Experiences of black americans," *HSR*, vol. 54, no. S2, pp. 1399–1408, 2019.

[4] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NeurIPS*, 2016, pp. 3323–3331.

[5] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," in *KAIS*, vol. 33, 2011, pp. 1–33.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ITCS*, 2012, pp. 214–226.

[7] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[8] F. Kamiran and T. Calders, "Classifying without discriminating," in *ICCCC*, 2009, pp. 1–6.

[9] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Interventional fairness: Causal database repair for algorithmic fairness," in *SIGMOD*, 2019, pp. 793–810.

[10] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *ICDE*, 2019, pp. 1334–1345.

[11] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, vol. 28, no. 3, 2013, pp. 325–333.

[12] A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You, "Training well-generalizing classifiers for fairness metrics and other data-dependent constraints," in *ICML*, vol. 97, 2019, pp. 1397–1405.

[13] H. Edwards and A. Storkey, "Censoring representations with an adversary," in *ICLR*, 2016.

[14] H. Zhao, A. Coston, T. Adel, and G. J. Gordon, "Conditional learning of fair representations," in *ICLR*, 2020.

[15] A. Singh and T. Joachims, "Policy learning for fairness in ranking," in *NeurIPS*, vol. 32, 2019, pp. 5427–5437.

[16] P. Awasthi, C. Cortes, Y. Mansour, and M. Mohri, "Beyond individual and group fairness," *arXiv preprint arXiv:2008.09490*, 2020.

[17] D. García-Soriano and F. Bonchi, "Maxmin-fair ranking: individual fairness under group-fairness constraints," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 436–446.

[18] 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.

[19] R. Binns, "On the apparent conflict between individual and group fairness," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 514–524.

[20] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *ICML*, vol. 80, 2018, pp. 3384–3393.

[21] S. Verma, M. Ernst, and R. Just, "Removing biased data to improve fairness and accuracy," *arXiv preprint arXiv:2102.03054*, 2021.

[22] C. Oh, H. Won, J. So, T. Kim, Y. Kim, H. Choi, and K. Song, "Learning fair representation via distributional contrastive disentanglement," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1295–1305.

[23] J. Liu, Z. Li, Y. Yao, F. Xu, X. Ma, M. Xu, and H. Tong, "Fair representation learning: An alternative to mutual information," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1088–1097.

[24] 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.

[25] Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi, "Understanding and improving fairness-accuracy trade-offs in multi-task learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1748–1757.

[26] C. Pinzón, C. Palamidessi, P. Piantanida, and F. Valencia, "On the impossibility of non-trivial accuracy in presence of fairness constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7993–8000.

[27] X. Li, P. Wu, and J. Su, "Accurate fairness: Improving individual fairness without trading accuracy," *arXiv preprint arXiv:2205.08704*, 2022.

[28] Anonymous, "Adversarial learning of group and individual fair representations (supplementary material)," 2023. [Online]. Available: https://github.com/anonym56492/E67196A1283BB901

[29] 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, pp. 2672–2680.

[30] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *ICCV*, pp. 2999–3007, 2017.

[31] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[32] H. Hofmann, "Statlog (German Credit Data)," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C5NC77.

[33] Agarwal, Sushant, "Trade-offs between fairness, interpretability, and privacy in machine learning," *UWSpace*, 2020.

APPENDIX A
EXTENDED DISCUSSION ON THE INCOMPATIBILITY BETWEEN GROUP AND INDIVIDUAL FAIRNESS

In this section, we first show the incompatibility of group fairness and individual fairness. Specifically, we 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. This motivates our goal to obtain a trade-off between group fairness and individual fairness.

As introduced in Section III-B, we use $\Delta yNN_{D,k}(\eta)$ to measure the level of individual fairness (i.e., how much the k-nearest neighbors of a data point x_i have consistent predicted outcome by classifier η as x_i for all $x_i \in D$). 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.

We consider the following question. *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 first introduce the concept of *k-NN cluster*. For any two data points $x_i, x_j \in D$, we connect them with an edge if $x_i \in \mathcal{N}_D^k(x_j)$ or $x_j \in \mathcal{N}_D^k(x_i)$. 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 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 $\mathcal{N}_D^k(x_j) \subseteq C, \forall x_j \in C$.

Now, consider a classifier η. If η satisfies both demographic parity and the yNN condition simultaneously, then it is easy to find that all the data points in the same k-NN cluster will be given the same prediction result (otherwise we should find a pair of similar data points with different labels, violating the yNN condition), which is a highly constrained condition. Moreover, to satisfy demographic parity, the positively predicted rates of all groups in the same k-NN cluster should also be equal, which makes this condition even more constrained. Another straightforward condition that satisfies both demographic parity and the yNN condition is that η gives the same predicted outcome to all data points, which is still a very restricted case [33]. We thus conclude that in most conditions, group fairness and individual fairness cannot be satisfied simultaneously. Besides, these constrained conditions are not desirable especially when we want to design an accurate classifier. Therefore, we show the incompatibility between the two kinds of fairness, and hence we should find an optimal trade-off between them.

APPENDIX B
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.

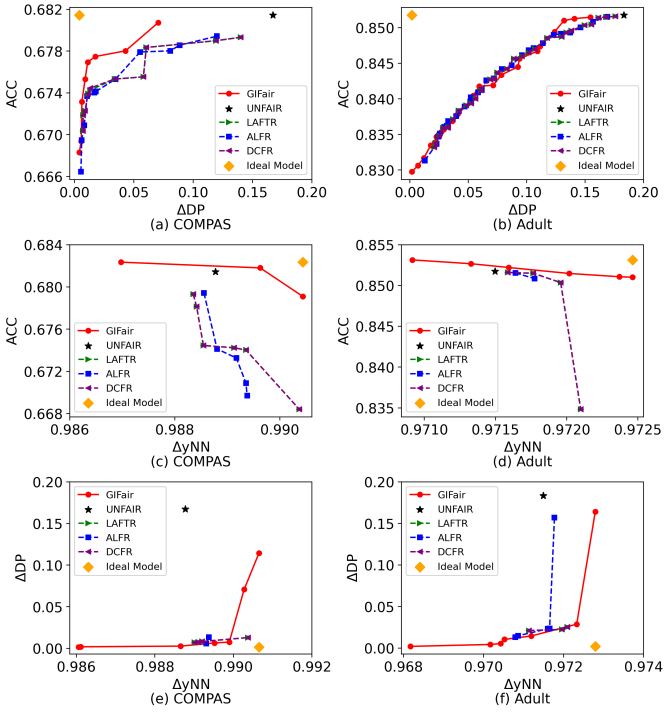

Fig. 4: Trade-off Curves on Dataset COMPAS and Adult

APPENDIX C
ADDITIONAL EXPERIMENTAL RESULTS

A. Remaining Trade-off Studies

On dataset COMPAS and Adult, we also observe the improved performance of GIFair over baselines on the trade-off between accuracy and group fairness (shown in Figure 4(a) and (b)), between accuracy and individual fairness (shown in Figure 4(c) and (d)) and between group fairness and individual fairness (shown in Figure 4(e) and (f)). For instance, on dataset COMPAS, GIFair dominates all other baselines in the range [0.005, 0.071] of ΔDP when trading-off accuracy and group fairness. Our dominating performance is also shown for trading-off accuracy and individual fairness on both dataset COMPAS and Adult, and for trading-off group fairness and individual fairness on dataset Adult.

B. 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

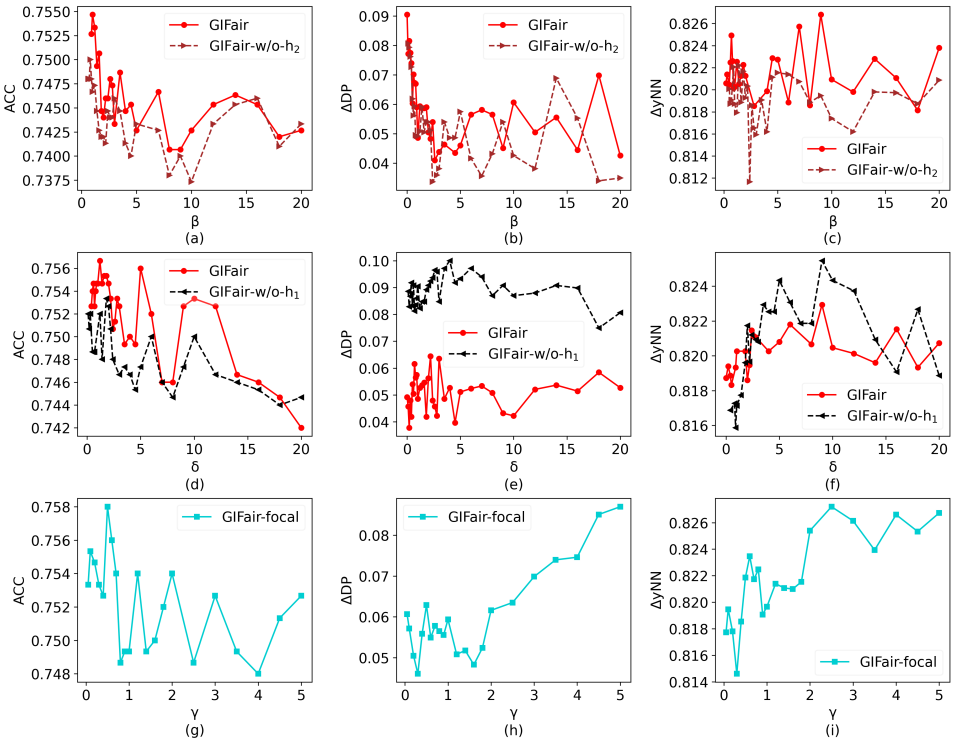

Fig. 5: Ablation Studies and Parameter Studies on Dataset German

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.

C. Remaining Ablation Studies and Parameter Studies

We present more results of our ablation studies and our parameter studies on all datasets as shown in Figure 5, 6 and 7.

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 5(a), (b), Figure 6(a), (b) and Figure 7(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 VI-E4) 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 7(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).

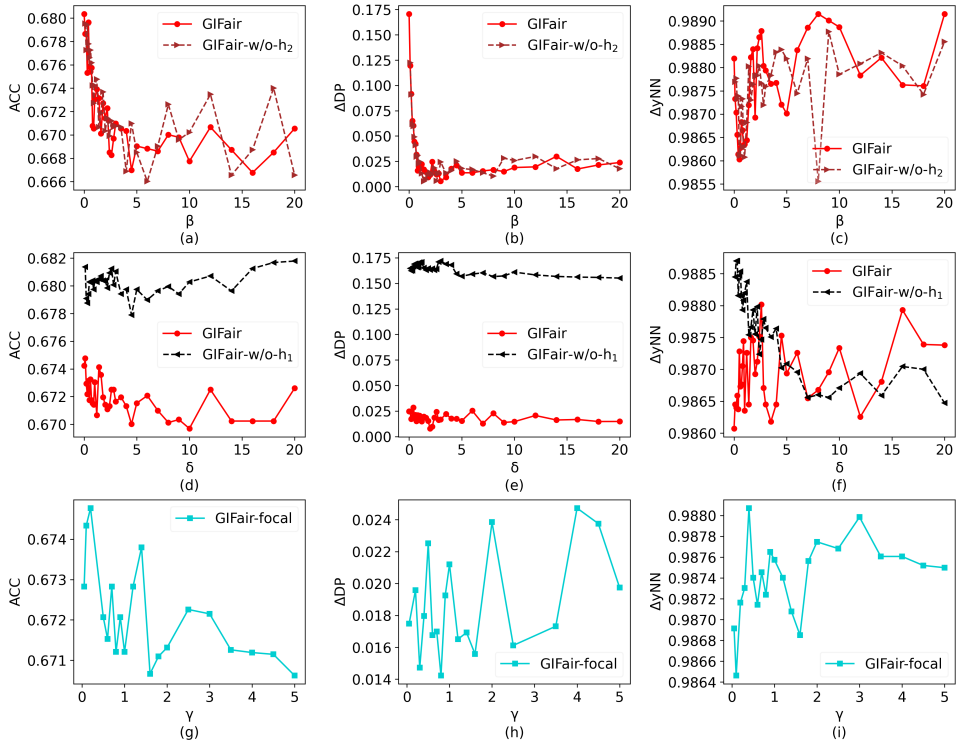

Fig. 6: Ablation Studies and Parameter Studies on Dataset COMPAS

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 5(d), (f), Figure 6(d), (f) and Figure 7(d), (f)). However, the change of δ does not obviously affect the performance of group fairness on dataset COMPAS (as shown in Figure 6(e)) and dataset Adult (as shown in Figure 7(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 5(g), (h), (i), Figure 6(g), (h), (i) and Figure 7(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.

Appendix D
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 3) and individual fairness (Equation 4). 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 $L_c(g(f(X)), Y)_D$, 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 $L_g(h_1(f(X)), A)_D$ defined on two groups), and then aggregate 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 $L'_g(h_1(f(X)), A)_D$, as follows.

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

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

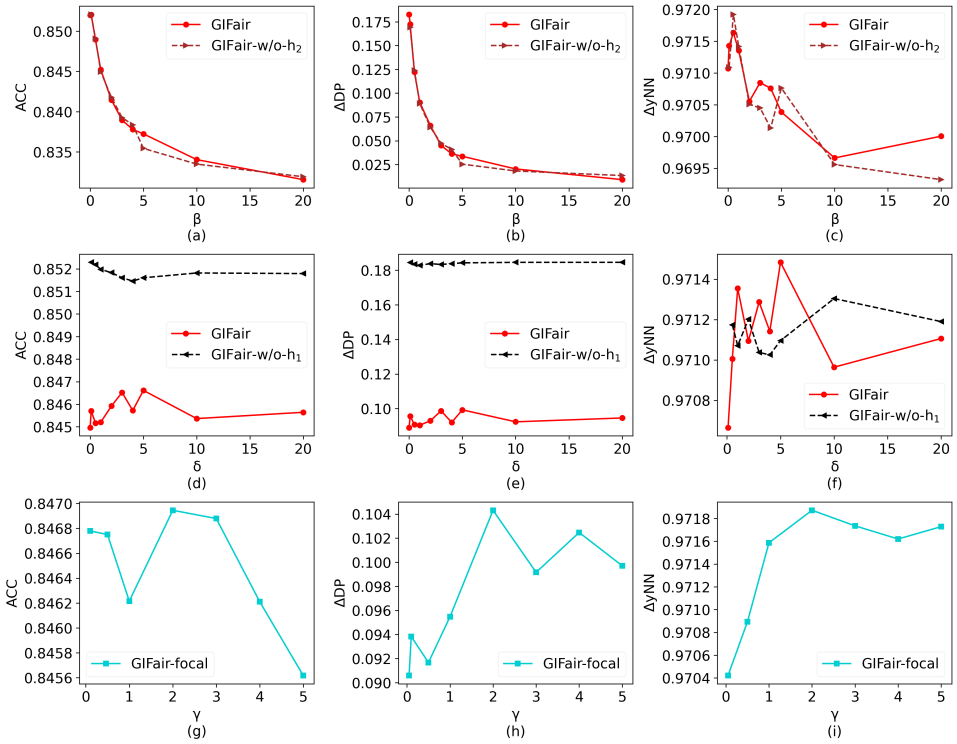

Fig. 7: Ablation Studies and Parameter Studies on Dataset Adult

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 $L'_g(h_1(f(X)), A)_D$. 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 $L_i(h_2(f(X)))_D$ to form a multi-outcome individual fairness loss function $L'_i(h_2(f(X)))_D$. Specifically,

$$L'_i(h_2(f(X)))_D = \frac{\sum_{x_i \in D} \sum_{x_j \in \mathcal{N}_D^k(x_i)} \mathcal{F}(h_2(f(x_i)) - h_2(f(x_j)))}{k \cdot N} \tag{10}$$

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 $L'_i(h_2(f(X)))_D$.

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.

Appendix E
Proof of Lemmas

Proof of Lemma IV.1. 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.

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

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

Proof of Lemma IV.2. It is easy to observe that $\Delta yNN'_{Z,k}(\eta)$ and $L_i(h^*_2(Z)_D)$ are formed similarly. Consider an individual adversary h'_2 that gives the same result as η, i.e., $h'_2 = \eta$. Then, $1 - \Delta yNN'_{Z,k}(\eta) = L_i(h'_2(Z)_D)$. Since the objective value of optimal individual adversary h^*_2 is no less than the objective value of any h'_2, we have $1 - \Delta yNN'_{Z,k}(\eta) \leq L_i(h^*_2(Z)_D)$. Clearly, $\Delta yNN'_{Z,k}(\eta) \geq 1 - L_i(h^*_2(Z)_D)$. □

APPENDIX