Individual Fairness with Group Awareness Under Uncertainty

Zichong Wang[1], Jocelyn Dzuong[1], Xiaoyong Yuan[2], Zhong Chen[3], Yanzhao Wu[1], Xin Yao[4], and Wenbin Zhang[1](✉)

[1] Florida International University, Miami, USA
{ziwang,wenbin.zhang}@fiu.edu
[2] Clemson University, Clemson, USA
[3] Southern Illinois University, Carbondale, USA
[4] Lingnan University, Hong Kong SAR, China

Abstract. As machine learning (ML) extends its influence across diverse societal realms, the need to ensure fairness within these systems has markedly increased, reflecting notable advancements in fairness research. However, most existing fairness studies exclusively optimize either individual fairness or group fairness, neglecting the potential impact on one aspect while enforcing the other. In addition, most of them operate under the assumption of having full access to class labels, a condition that often proves impractical in real-world applications due to censorship. This paper delves into the concept of individual fairness amidst censorship and also with group awareness. We argue that this setup provides a more realistic understanding of fairness that aligns with real-world scenarios. Through experiments conducted on four real-world datasets with socially sensitive concerns and censorship, we demonstrate that our proposed approach not only outperforms state-of-the-art methods in terms of fairness but also maintains a competitive level of predictive performance.

Keywords: Censorship · Group fairness · Individual fairness

1 Introduction

In recent years, the widespread utilization of machine learning algorithms in various domains has raised a growing societal concern regarding the bias and discrimination inherent in these algorithms. This is particularly consequential in high-stakes decision-making scenarios such as job applicant ranking [8], criminal justice [11], and credit scoring [51]. To this end, algorithmic fairness has garnered significant attention, leading to the development of a large collection of fair ML notions and algorithms [42]. These are typically categorized as either group or individual fairness [33]: Group fairness involves identifying *sensitive attributes* (*e.g.,* race or gender) which could be potential sources of bias, and then evaluating whether the outcome statistics of the classifiers (*e.g.,* prediction accuracy

A. Bifet et al. (Eds.): ECML PKDD 2024, LNAI 14945, pp. 89–106, 2024.
https://doi.org/10.1007/978-3-031-70362-1_6

and true positive rate) are similar across different subgroups [39]; Individual fairness studies bias at a much finer granularity, ensuring that similar individuals receive similar probability distributions over class labels, thereby mitigating unfair treatment [19].

The majority of these prior studies approach fairness as a supervised learning problem, presuming the presence of class labels to quantify and mitigate bias. However, this assumption is unrealistic in many real-world applications due to the prevailing censorship [22,28,41]. For instance, consider Fig. 1 as an example of an ML-based job screening task. Here, the actual application results, *i.e.,* class labels, remain unknown for censored individuals like d_3 and d_4. This censoring phenomenon can be attributed to various causes. For instance, in the case of applicant d_4, the study concluded before the applicant received their application result, resulting in a lack of information about d_4's class label. In other cases, applicants may become lost to follow-up, withdraw, or experience competing events that make further follow-up impossible. For instance, the applicant d_3 might have received alternative job opportunities and chosen to withdraw from the application system, leading to an unknown application outcome or class label. Due to the inability to handle such censorship information, existing fairness works either exclude observations with uncertain class labels [11,18,52] or ignore the censorship information [38,40]. Both strategies can introduce substantial bias, as censored information contains important details and cannot simply be ignored [6]. For instance, this omission can skew critical components of individual fairness, such as the similarity metric, leading to inappropriate similarity evaluation due to the exclusion of censorship information [5].

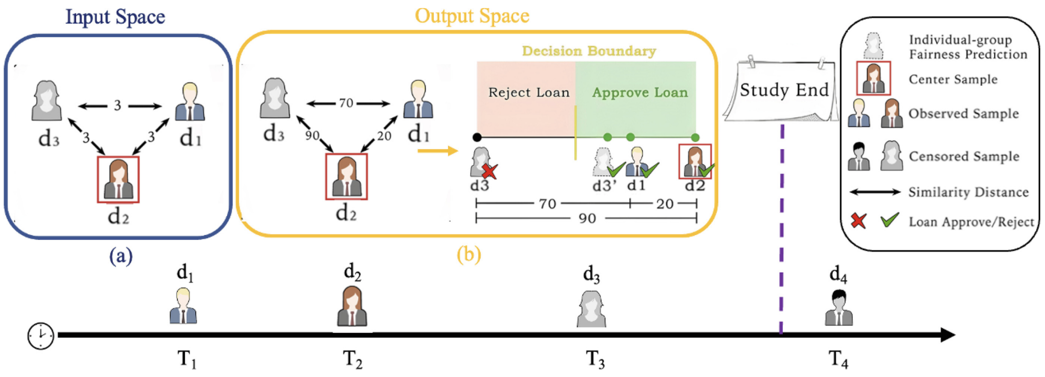


Fig. 1. An example depicts the issue of censoring and disparity in individual fairness without group awareness. Individuals d_3 and d_4 are censored, while others, *i.e.,* d_1 and d_2, are not censored. Individuals are arranged in ascending order of their survival times, with the shortest survival time, *i.e.,* T_1, at the far left. The study concludes at the orange dashed line and the center sample provides prior information to predict other samples. (Color figure online)

Moreover, existing research on fairness often treats individual fairness and group fairness as distinct tasks, failing to consider potential implications among them [19,26,55]. However, this separation of objectives could introduce additional bias into each other. Still considering the example depicted in Fig. 1, the centered individual d_2 has her loan application known as approved while the loan decisions for male applicant d_1 and female applicant d_3 will be determined by the individual fairness enforced ML algorithm. As we can see in the input space (Fig. 1 (a)), these individuals have equal similarity distances: (d_1,d_2) = (d_1,d_3) = (d_2,d_3)=3, while the distances become (d_1,d_2) = 20, (d_1,d_3) = 70, and (d_2,d_3) = 90 in the output space (Fig. 1 (b)). This suggests these three individuals are alike with respect to their input attributes. However, in the output space, while the results are individual fairness constraints enforced, d_1 continues to exhibit similarity to d_2, whereas d_3 becomes distinct from d_2. This disparity arises as the average constraint scalar is inconsistent across groups for individual fairness constraints without group awareness (*i.e.,* the average constraint scalar for males is higher than that for females); d_1 is under more stringent individual fairness constraints, positioning him nearer to d_2, while d_3, experiencing more lenient individual fairness constraints, is positioned farther from d_2. As a result, d_1, who is closer to d_2 in the output space, receives the same approval outcome, while d_3, who is farther away, receives a distinct decline decision, leading to an inequitable outcome for her job application (d_3, with group awareness, would be placed to the position of d'_3, ensuring a fair outcome). In addition, while metrics evaluating both input and output space similarities can be properly defined by domain experts, the Lipschitz condition needed in existing individual fairness studies to calibrate the distance between them is non-trivial to specify, which has been another major obstacle to wider adoption in real-world applications [34].

Therefore, there is a pressing need to address fairness in the presence of censorship while simultaneously balancing the impact of individual and group fairness, which is still largely unexplored and presents unique challenges: **i) Quantifying and mitigating bias in censored settings.** Most of the existing fairness notions and algorithms often overlook censorship information, rendering them inapplicable [2,4,39]. **ii) Achieve a balance between individual and group fairness.** Most existing fairness work considers individual or group fairness independently, overlooking the potential complications between them [37]. **iii) Unbounding the Lipschitz condition.** Existing individual fairness works usually rely on the Lipschitz condition to align the difference in the metrics of input and output spaces, restricting its applicability in real-world scenarios [34].

To tackle the above challenges, this paper investigates fairness with censorship and the interplay between group and individual fairness for fairness guarantees more in line with realistic assumptions. The key contributions of this paper can be summarized as follows:

- We present a new research challenge that focuses on the intersection of individual and group fairness in learning with censorship.
- We analyze the impact of individual fairness constraints on group fairness, and devise notions to measure individual and individual-group bias amidst censorship.

- We introduce fairCox, an individual-centric debiasing algorithm with built-in group awareness, custom-designed for applications amidst censorship.
- Extensive experimental results on four real-world socially sensitive datasets with censorship.

The organization of this paper is as follows: A review of the pertinent literature is provided in Sect. 2. Section 3 introduces the notations and problem definition. Our proposed approach and fairness metrics are thoroughly explicated in Sect. 4. Section 5 elucidates the experimental framework and discusses the outcomes. Finally, Sect. 6 concludes the paper.

2 Related Work

In this section, we give a brief overview of related work on fairness in ML and survival analysis.

2.1 Fairness in ML

Fairness in ML is a well-explored research area with numerous studies conducted to quantify and mitigate ML bias [12,13,43,45,46,56]. Existing fairness work can be typically categorized into group and individual fairness [31]. The former [16,32,44,47,48,50] seeks to ensure statistical equality among subgroups defined by sensitive attribute(s), while the latter [3,19,21,35,37] aims to ensure that similar individuals receive similar probability distributions over class labels, promoting equitable treatment irrespective of individuals' sensitive attributes. While individual fairness offers a finer granularity in scrutinizing potential biases and discrimination compared to group fairness, it relies on the Lipschitz condition [3]. Specifically, individual fairness requires that the similarity distance between individuals in the output space should not exceed their similarity distance in the input space [19]. However, specifying a suitable Lipschitz constant to compare these distances accurately can be difficult, due to the variation in distance metrics between the input and output spaces. Moreover, most existing fairness works often prioritize a single fairness goal-be it individual or group fairness-overlooking the potential implications among various fairness objectives. As discussed in Sect. 1, this oversight has the potential to introduce additional bias, leading to deprived subgroups facing persistent challenges in accessing loans [36]. Additionally, a common limitation of these approaches is the inherent assumption of the availability of class labels, rendering them inapplicable in censorship settings where labels can be uncertain [54].

2.2 Survival Analysis

The prevalence of survival data, also known as censored data, motivates the study of *survival analysis* to address the challenges of accessing partial survival information from study cohorts [14]. Among the numerous survival models proposed, the Cox Proportional Hazards (CPH) model stands out as the most widely utilized [15], describing the multiplicative relationship between risk, as conveyed by the hazard function, and covariates. In recent times, deep neural networks have also been employed to capture the nonlinearity of censored data [25]. In contrast, an alternative approach explores tree-based methods [24], with a particular focus on random forests due to their superior capabilities in handling the nonlinear effects of variables and free of assumptions like proportional hazards. Given the widespread use of survival models, it becomes imperative to include considerations of fairness. Recent work in this domain includes FSRF [55], which aims to achieve group fairness amidst censorship by ensuring consistency in pairwise comparisons between model predictions and true outcomes across subgroups. On the other hand, IDCPH [26] shifted focus towards individual fairness in contexts with censored data. However, IDCPH's reliance on the Lipschitz condition can limit its applicability in various contexts. To this end, IFS [53] attains censored individual fairness by enforcing consistency based on rankings between the input and output spaces. A significant drawback, however, lies in its surrogate assessment of individual fairness loss. Specifically, IFS quantifies bias based on the focused individual to form the ranking, using the bias surrogate, instead of considering individuals within the ranking who are the actual subjects of discrimination.

To jointly address these challenges, our method proposes a holistic approach that aims for individual fairness in censored settings while concurrently ensuring equitable treatment across various subgroups, and it also bypasses the Lipschitz constraints present in traditional individual fairness methods.

3 Notations and Problem Definition

Each individual d_i in the censored data D can typically be described as $d_i = \{x_i,$ T, $\delta\}$, where: i) x_i denotes the set of observed features with a special attribute, referred as sensitive attribute s_i, that differentiates between favored and deprived subgroups (*e.g.,* male vs. female), ii) T denotes the survival time, *i.e.,* the time of the event, and iii) δ is the event indicator, signaling whether the event has been observed (*i.e.,* δ=1 when the event is observed otherwise the survival time is censored, creating uncertainty to the class label).

To model survival data, the *hazard function* is commonly used, which stipulates the instantaneous rate of event occurrence at a specified time t:

$$h(t|x) = \lim_{\triangle t \to 0} \frac{\mathrm{Pr}\ (\mathrm{t} < \mathrm{T} < \mathrm{t} + \triangle \mathrm{t} | \mathrm{T} \geq \mathrm{t}, \mathrm{x})}{\triangle t} \tag{1}$$

Given the hazard function, we can compute the *survival function*, $S(t|x) =$ Pr $(T > t|x)$, which indicates the probability that an event occurs after a specific time t, and vice versa. Mathematically, as shown in Eq. 2:

$$S(t|x) = \exp\left(-\int_0^t h(t|x)dt\right),\ h(t|x) = h_0(t)\exp(\beta^\top x). \tag{2}$$

where $h_0(t)$ represents the baseline hazard function (*i.e.,* when $x = 0$). Additionally, β is a set of undetermined parameters that can be estimated as the partial likelihood in Eq. 3:

$$L(\beta) = \prod_{i:\delta_i=1} \frac{\exp(\beta^\top x_i)}{\sum_{j:T_j \geq T_i} \exp(\beta^\top x_j)}, \ \hat{\beta} = \arg\max_{\beta} L(\beta) \tag{3}$$

In this work, our primary goal is to quantify individual unfairness with group awareness amidst censorship as well as to model a fair survival function $H(\cdot)$ to mitigate the quantified bias amidst censorship.

4 Methodology

The presence of censorship restricts the suitability of widely employed fairness notions and algorithms that have been introduced in existing ML fairness studies. In addition, existing works often treat individual and group fairness as separate goals, resulting in models that enhance individual fairness at the cost of group fairness. To this end, our proposed model aims to measure individual fairness amidst censorship precisely. By pinpointing the origins of individual bias, it also becomes feasible to further quantify associated group-level bias. Specifically, in Sect. 4.1, we start by outlining a criterion to identify the origins of individual bias from a ranking perspective. The intuition is that individuals should have the same position in the input space and output space obtained through different reference individuals. This enables a direct evaluation of individuals facing unfair treatment, while also facilitating the measurement of group-level bias. Expanding upon this, Sect. 4.2 further introduces a metric specifically designed to quantify the disparity of individual unfairness between groups. Finally, Sect. 4.3 presents the model to achieve individual fairness with group awareness under censorship.

4.1 Quantifying Individual Fairness with Censorship

This section introduces a novel individual fairness notion, namely *Cumulative Ranked Individual Fairness (CRIF@k)*, specifically tailored to address censorship scenarios, offering a refined approach to evaluating bias by examining individuals within the ranking list instead of the reference individual to form the list. This is achieved by assessing both the direct discriminatory effect on the individuals themselves and the indirect impact through their neighbors. This strategy can identify individuals who are genuinely discriminated against, while at the same time circumventing the constraints imposed by the Lipschitz condition.

To illustrate the overarching concept, take the example shown in Fig. 2. The rankings, established based on similarity with d_1 as the reference, are organized in descending order in both the input and output spaces, represented as $\{d_2, d_3, d_4\}$ and $\{d_2, d_4, d_3\}$, respectively. In this example, the bias originates within the ranking list (*i.e.,* d_2, d_3, d_4) rather than the reference d_1. To quantify such ranking-based bias, we evaluate the changes in the ranking of each d_i (*i.e.,*

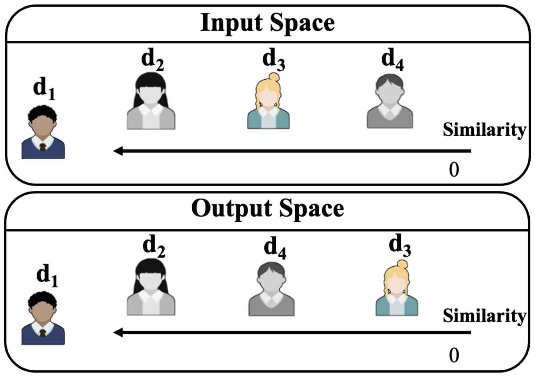


Fig. 2. An illustration of quantifying individual bias from a ranking perspective, where similarity rises from 0 moving from right to left.

d_2, d_3 or d_4) relative to the reference d_r (*i.e.,* d_1) and the shift in the relative positions of d_i and his/her neighbors, which also simultaneously eliminates the need for specifying a Lipschitz constant, thus improving the applicability. Taking d_3 as a specific example, let's illustrate two types of associated bias. First, there is a discrepancy in his/her proximity to the reference d_1 in the output space compared to the input space, potentially receiving different outcomes of their socially sensitive application. Second, a relative disadvantage compared to his/her neighbor d_4 in the output space. This dual consideration allows for a more nuanced understanding of how an individual's ranking position is affected in different contexts. Finally, the overall individual unfairness of d_3 is determined by calculating the average value of d_3's individual unfairness across all his/her associated ranking lists. Mathematically, the proposed CRIF is defined as follows:

$$\text{CRIF@}k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{r=1}^{M}\frac{\text{RIF}^{d_r}_{\text{Sim}_{D'}(d_i)}}{\text{RIF}^{d_r}_{\text{Sim}_{D}(d_i)}} \tag{4}$$

where N is the total number of individuals, k refers to the length of the top-k ranking list formed with d_r as the reference, and M indicates the number of occurrences where d_i is among these top-k ranking lists (emphasizing the top-k individuals in line with the fundamental principle of individual fairness, which mandates that similar individuals should be treated similarly). In addition, D' and D indicate the output and input spaces, respectively, while the formulation $\text{RIF}^{d_r}_{\text{Sim}_{(\cdot)}(d_i)}$ denotes the *Ranked Individual Fairness* of d_i,

$$\text{RIF}^{d_r}_{\text{Sim}_{(\cdot)}(d_i)} = \frac{\text{Sim}_D(d_i)}{\log_2(\text{pos}(\ d_i\ )+1)} + \frac{1}{k-1}\sum_{\substack{\text{pos}=1\\ d_{l_{\text{pos}}}\neq d_i}}^{k}\frac{\text{Sim}_D(d_{l_{\text{pos}}}, d_r)}{\log_2(\text{pos}+1)} \tag{5}$$

where $(\cdot)$ can be either D' or D, sequence $\{l_{\text{pos}}\}_{\text{pos}=1}^{k}$ represents the ordered list of individual indices obtained from the similarity matrix $Sim(\cdot)$ for the reference

d_r, and $Sim_D(d_{l_{\text{pos}}}, d_r)$ denotes the similarity in the input space between the individual at the *pos*-th position of the ordering list, $d_{l_{\text{pos}}}$, and the reference d_r. Note that both $\text{RIF}_{\text{Sim}_D(d_i)}$ and $\text{RIF}_{\text{Sim}_{D'}(d_i)}$ compute the RIF using the similarity values from Sim_D (*i.e.,* input space), with the corresponding similarity being used only for deriving the ordering list l_{pos}, which directly quantifies individual bias while eliminating the necessity of specifying a Lipschitz constant. In addition, Sim_D is typically contingent upon the specific problem at hand and can be assessed by domain experts [29,30], while the exponential of the negative difference in risk scores is utilized as the similarity metric for $Sim_{D'}$, and $Sim_{D'}$ can thus be formally defined as:

$$\text{Sim}_{D'(d_r,d_i)} = (1 - (|C_{x_r} - C_{x_i}|)) \times \exp\left(-|\exp(\beta^\top x_r) - \exp(\beta^\top x_i)|\right) \qquad (6)$$

where the feature x_i characterizes individual d_i (x_i and d_i are used interchangeably for simplicity). The term $|C_{x_r} - C_{x_i}|$ measures the concordance difference between d_r and d_i, which adjusts the similarity between these two individuals, while incorporating essential survival information. This adjustment is necessary and achieved by measuring the difference in the concordances between two individuals, ensuring that the similarity measure accurately reflects the orders among all individuals that can actually be ranked. Specifically, the concordance C_{x_i} for an individual d_i is defined as follows:

$$C_{x_i} = \frac{1}{\sum_{x_j \neq x_i} \mathbb{1}[\delta_< = 1]} \sum_{x_i \neq x_j} \mathbb{1}[h(t|x_>) < h(t|x_<)|\delta_< = 1] \qquad (7)$$

where $\mathbb{1}$ represents the indicator function, and $x_>$ and $x_<$ denote individuals with longer ($T_> = \max(T_{x_i}, T_{x_j})$) and shorter survival times ($T_< = \min(T_{x_i}, T_{x_j})$) respectively. The event indicator of the shorter survival time ($T_<$) is denoted as $\delta_<$; $\delta_<$=1 means shorter time is not censored, and thus this pair-wise comparison is permissible for comparison. With identified permissible pairs, C_{x_i} represents the proportion of all other individuals that are correctly ranked with the sample x_i in relation to their actual survival times.

Overall, our proposed CRIF@k is obtained by calculating the average ratio between the RIF from the output space reflected in the input space and the actual input space. Therefore, as the CRIF@k score increases, ranging between 0 and 1, the model becomes fairer due to enhanced consistency between orderings in its input and output spaces. In other words, if two individuals are ranked closely in the input space (*e.g.,* their personal circumstances), then they should also be ranked closely in the output space (*e.g.,* their application results).

4.2 Quantifying Individual-Group Fairness with Censorship

To further tackle the limitation of neglecting the implications between individual and group fairness in existing fairness works, this section introduces a novel metric to assess the disparity of individual fairness among various subgroups, as well as in the presence of censorship. To accomplish this goal while adhering to the core principle of group fairness, we employ CRIF, as explained in the preceding section, to evaluate if the ML model treats different subgroups equally when assigning favorable outcomes. Specifically, we evaluate the individual unfairness of each individual d_i within a subgroup $\mathbf{D}_{s_i}$, thereby evaluating the model's parity across each subgroup. Building on this, *Group Fairness (GF)*, which is designed to measure the level of individual fairness of each subgroup $\mathbf{D}_{s_i}$ is formally defined in Eq. 8:

$$\mathrm{GF}_{\mathrm{D}_{\mathrm{s_i}}} = \frac{1}{|\mathrm{D}_{\mathrm{s_i}}|} \sum_{\mathrm{i=1}}^{|\mathrm{D}_{\mathrm{s_i}}|} \frac{1}{\mathrm{M}} \sum_{\mathrm{r=1}}^{\mathrm{M}} \frac{\mathrm{RIF}^{\mathrm{d_r}}_{\mathrm{Sim}_{\mathrm{D'}}(\mathrm{d_i})}}{\mathrm{RIF}^{\mathrm{d_r}}_{\mathrm{Sim}_{\mathrm{D}}(\mathrm{d_i})}} \tag{8}$$

where $|\cdot|$ represent the number of individuals in the subgroup D_{s_i}. Based on $\mathrm{GF}_{\mathrm{D}_{\mathrm{s_i}}}$, the *Individual Group Disparity (IGD)* is introduced to capture the largest deviation in the model's discriminative abilities across different demographic groups D_{s_i} and D_{s_j} during the optimization process for individual fairness, and is mathematically defined as follows:

$$\mathrm{IGD} = \min_{\forall\ \mathbf{D}_{s_i}, \mathbf{D}_{s_j} \in \mathbf{D}, \mathbf{D}_{s_i} \neq \mathbf{D}_{s_j}} \left\{ \left| \frac{\mathrm{Min}(\mathrm{GF}_{\mathbf{D}_{s_i}}, \mathrm{GF}_{\mathbf{D}_{s_j}})}{\mathrm{Max}(\mathrm{GF}_{\mathbf{D}_{s_i}}, \mathrm{GF}_{\mathbf{D}_{s_j}})} \right| \right\} \tag{9}$$

The $\mathrm{IGD}_{\mathbf{D_{s_i}}}$ value ranges from 0 to 1, with a score of 1 indicating equal individual fairness among all subgroups, signifying unbiased treatment, while a score of 0 symbolizes a disparity in treatment between subgroups, indicating biased outcomes. To summarize, utilizing IGD allows for the identification of any disparity across groups, revealing whether the model favors certain subgroups over others while enforcing individual fairness amidst censorship.

4.3 Mitigating Bias Under Censorship

With the integration of fairness definitions that meticulously account for censorship, we introduce a debiasing algorithm, *fairCox*. This is structured around the standard Cox proportional hazard model, devised to produce forecasts that ensure equitable risk predictions across individuals. In addition, *fairCox* simultaneously harmonizes the disparities in individual unfairness amongst varying subgroups. Essentially, *fairCox* augments the partial likelihood maximization of the CPH model by integrating our individual and individual-group fairness quantification metrics, CRIF@k and IGD. Its design is such that it aspires to individual fairness while diminishing the disparity in individual fairness across different subgroups. Below, we detail the loss functions for each optimization objective and present the total objective function for optimizing our framework.

First, to maintain the maximization of model utility, the utility loss function $\mathcal{L}_{\text{utility}}$ is formulated as the negative log partial likelihood of the CPH model.

Based on the partial likelihood presented show in Eq. 3, we define $\mathcal{L}_{\text{utility}}$ as follows:

$$\mathcal{L}_{\text{utility}} = - \sum_{i:\delta_i=1} (\beta^\top x_i - \log \sum_{j:T_j \geq T_i} \exp(\beta^\top x_j)) \tag{10}$$

Second, we incorporate individual fairness quantification as the individual fairness regularizer, denoted as $\mathcal{L}_{\text{if}} = \text{CRIF}@k$, along with the loss function, denoted as L_{gf}, for the IGD objective, aiming to promote group equality with respect to individual fairness. Specifically, our aim is to equalize the levels of individual unfairness across all subgroups, where $\mathcal{L}_{\text{igf}} = \text{IGD}$. Please note that this loss function exhibits symmetry between any two given groups. Finally, we define the unified objective function as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{utility}} + \alpha \mathcal{L}_{\text{if}} + \mu \mathcal{L}_{\text{igf}} \tag{11}$$

In summary, *fairCox*, weighted by the hyperparameters α and μ, jointly optimizes utility, individual fairness, and group equality-informed individual fairness objectives for individuals with group-awareness fairness learning amidst censorship.

5 Experiment

5.1 Datasets

We validate the proposed model using four real-world censored datasets that involve socially sensitive concerns and exhibit a range of distinct characteristics: i) The *ROSSI* dataset comprises data on individuals who were convicted and subsequently released from Maryland state prisons, with a one-year follow-up period [22]. ii) The *COMPAS* dataset, a pivotal resource in algorithmic fairness research, contains information used to predict recidivism rates in Broward County [1]. iii) The *KKBox* dataset is derived from the WSDM-KKBox's Churn Prediction Challenge 2017 [28]. iv) The *Support* dataset offers data on patients admitted to five tertiary care academic centers [27]. Table 1 presents a summary of key details for these datasets, with a notable emphasis on the explicit inclusion of survival information, aligning with real-world scenarios.

5.2 Baselines

To assess the effectiveness of our approach, we compared it against six baseline models: IDCPH [26], FSRF [55], CPH [15], RSF [24], DeepSurv [25], and IFS [53]. Among them, IDCPH and IFS have recently introduced a censored individual fairness model, while FSRF focuses on group-level fairness considerations. In addition, CPH is the conventional and widely utilized approach to model censored data, RSF is considered a state-of-the-art survival model using random forests, and DeepSurv is a recent deep learning model designed for censored data. Other fairness methodologies are not included for comparison as they are incapable of handling censorship information by design.

Table 1. Summary of the datasets used in the evaluations.

	ROSSI	COMPAS	KKBox	Support
Sample#	432	10,325	2,814,735	8,873
Feature#	9	14	18	14
Sensitive Attribute	Race	Race	Gender	Gender
Sensitive Value	African American	African American	Female	Female
Censored#	318	7,558	975,834	2,840
Censored Rate%	73.6	73.2	34.7	32.0

5.3 Evaluation Metrics

Our evaluation encompasses a range of fairness and performance metrics to provide a thorough assessment. To gauge fairness, the proposed individual fairness metrics CRIF@10 and individual-group fairness metrics IGD are employed. Note that existing widely-used fairness metrics could not be applied as they are not adaptable to censorship settings. In evaluating performance, we align with prior works [26,55] and consider three typical survival performance metrics: C-index, Brier score, and Time-dependent AUC. The C-index, introduced by [23], assesses a model's ability to discriminate between outcomes through the correct pairwise ordering and equals to the area under ROC Curve (AUC) in the absence of censorship. The Brier score [7], on the other hand, calculates the mean squared discrepancy between the predicted probability of outcome assignment and the actual outcome. Finally, the Time-dependent AUC [9] quantifies the probability that a randomly chosen pair of individuals, one who has experienced the event and another who hasn't at time t, are correctly ranked. A higher value is preferred for both C-index and Time-dependent AUC, whereas a lower Brier score is indicative of better prediction. To demonstrate the generalization of fairCox, we build $\text{Sim}_{D'}$ by incorporating the Euclidean distance with feature scaling.

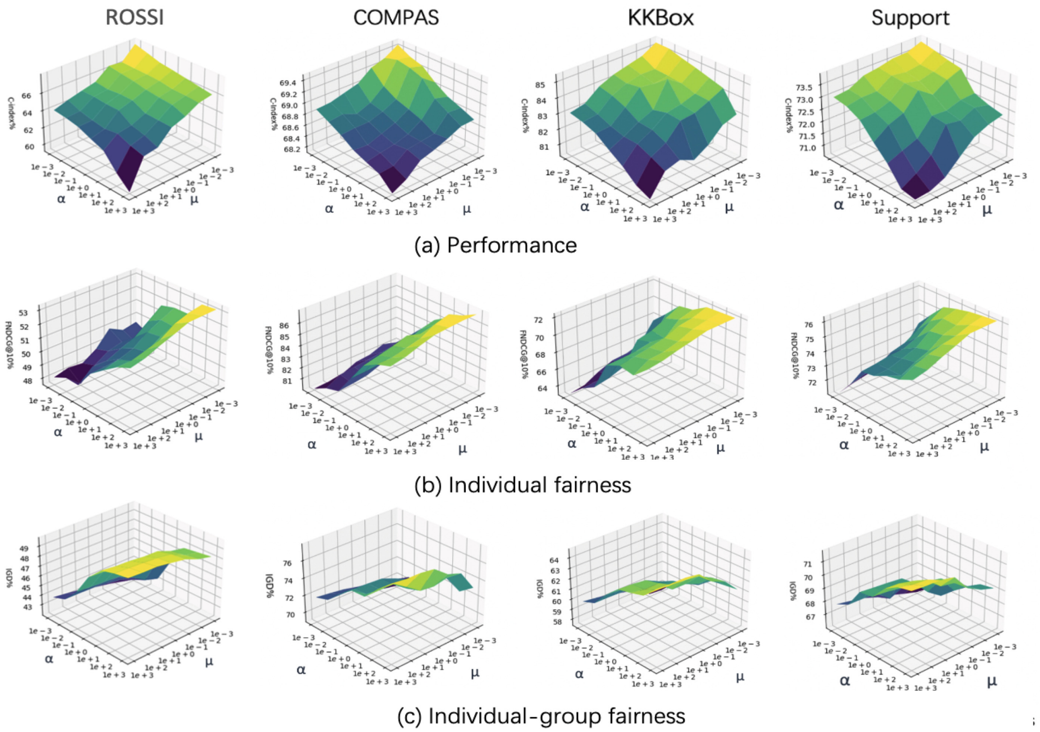
5.4 Experimental Results

Evaluation of the Performance and Fairness of fairCox. To evaluate the effectiveness of fairCox, we compare its performance and fairness against six state-of-the-art baselines. All methods are trained using the same procedure to ensure a fair comparison, and the results from 5-fold cross-validation are summarized in Table 2. As we can see, fairCox significantly outperforms all other baselines in reducing discrimination (CRIF@10 and IGD) while achieving comparable prediction performance, measured by C-index, Brier score, and Time-dependent AUC. Specifically, baseline methods exhibit inferior fairness performance, attributable to either a disregard for fairness considerations or an overlooked intermediate potential relationship between individual and group fairness. Even against the FSRF baseline, which is specifically for group fairness,

Table 2. Evaluation results of different models with the best results marked in bold. The numbers in parentheses represent the relative performance improvement of fairCox compared to the best baseline (Bolding indicates the best results).

Dataset	Method	Metrics				
		CRIF@10%	IGD%	C-index%	Brier Score%	Time-dependent AUC%
ROSSI	IDCPH	43.41	32.77	52.28	25.68	63.92
	FSRF	26.82	20.53	61.44	14.66	65.12
	CPH	33.41	24.65	64.24	19.45	65.46
	RSF	36.17	29.39	65.47	15.05	79.54
	DeepSurv	31.43	23.15	**66.67**	**14.52**	**80.12**
	IFS	45.83	30.78	65.78	14.79	77.63
	fairCox	**50.63**	**48.94**	63.67	15.07	78.35
		(10.47%)	**(49.34%)**	(−4.49%)	(−3.79%)	(−2.21%)
COMPAS	IDCPH	70.27	62.18	62.16	23.37	60.30
	FSRF	40.41	31.38	52.28	13.78	63.92
	CPH	73.51	59.73	69.24	18.89	67.72
	RSF	72.64	62.47	72.61	13.02	71.33
	DeepSurv	73.78	63.85	**75.21**	**12.54**	**73.68**
	IFS	74.27	63.47	73.83	12.98	71.47
	fairCox	**77.65**	**74.91**	71.47	12.83	70.67
		(4.55%)	**(17.32%)**	(−4.97%)	(−2.31%)	(−4.01%)
KKBox	IDCPH	56.61	51.72	72.61	21.13	73.31
	FSRF	38.75	56.85	78,53	**13.57**	79.72
	CPH	47.32	44.63	80.02	17.42	78,47
	RSF	42.41	38.69	82.32	13.84	80.22
	DeepSurv	43.45	39.54	**83.01**	14.32	80.69
	IFS	57.60	54.78	81.97	14.41	80.04
	fairCox	**64.47**	**62.85**	82.43	14.45	**81.95**
		(11.93%)	**(14.73%)**	(−0.69%)	(−6.48%)	(1.56%)
Support	IDCPH	62.53	53.14	62.58	28.53	72.72
	FSRF	50.15	62.47	59.28	**12.98**	73.92
	CPH	58.92	49.82	69.31	20.31	77.64
	RSF	51.17	44.57	71.73	15.50	80.77
	DeepSurv	53.44	46.30	**72.32**	14.89	81.13
	IFS	65.61	62.82	70.03	15.37	**81.38**
	fairCox	**72.17**	**70.40**	70.69	13.83	78.47
		(14.88%)	**(12.07%)**	(−2.25%)	(−6.55%)	(−3.56%)

fairCox yields remarkable results in promoting group fairness. This is particularly noteworthy since FSRF, despite its focus on group fairness, fails to address the dynamic interplay between individual and group fairness, leading to outcomes that may not be fair to every individual within the group. This finding underscores the importance of incorporating both individual and group fairness considerations. By integrating considerations for both individual and group fairness, fairCox distinctly overshadows baselines that focus solely on one, illustrating its comprehensive advantage in fostering fairness. Additionally, the enhancement in the overall predictive performance of fairCox underlines the importance of antidiscriminatory designs in improving prediction accuracy, presumably due to the reduction of overfitting through fairness regularization.



(a) Performance

(b) Individual fairness

(c) Individual-group fairness

Fig. 3. Exploring hyperparameters study results in four real-world datasets.

Effect of Different α and μ Values on Fairness and Predictive Performance. Two critical hyperparameters, α and μ, are employed to optimize individual fairness and individual-group fairness objectives of fairCox, respectively. To evaluate the effects of them on both the performance and fairness, we conduct hyperparameter sensitivity experiments, varying α and μ within the set $\{1e^{-3}, 1e^{-2}, 1e^{-1}, 1e^{0}, 1e^{1}, 1e^{2}, 1e^{3}\}$. Figure 3 presents the results of the hyperparameter sensitivity analysis. As one can see, it is evident that increments in α and μ are inversely proportional to the performance of fairCox. Moreover, an increase in α enhances the model's individual fairness performance and slightly moderates the discrepancy in individual fairness performance across diverse subgroups. This occurs as all samples are subjected to stringent constraints, however, it will lead to a significant decrease in performance. Conversely, elevating μ allows the model to equilibrate the individual fairness performance variances between different subgroups more effectively.

Effect of Different Numbers of Neighbors k-values on Individual Fairness and Predictive Performance. In our framework, different values of the number of neighbors k affect model fairness and predictive performance. Hence,

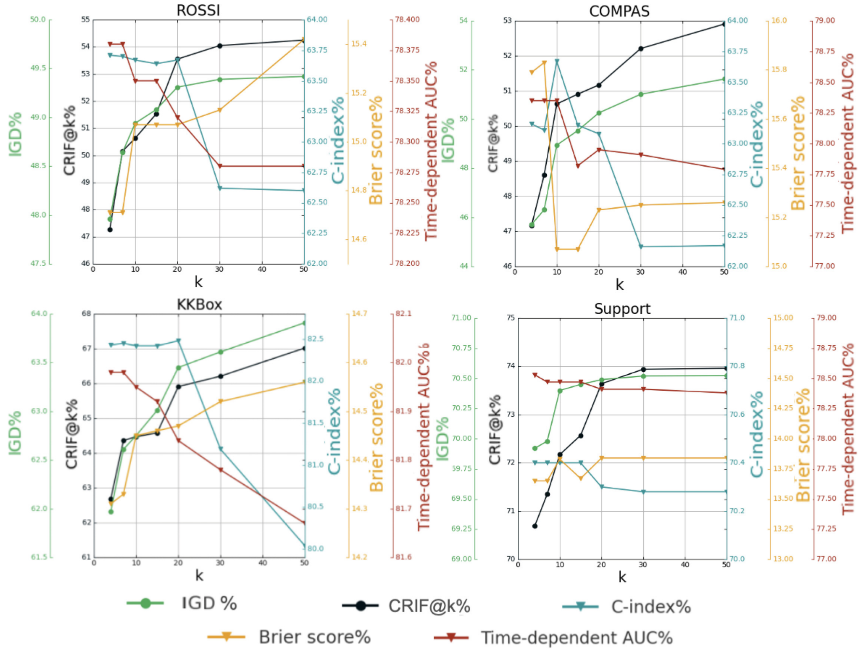


Fig. 4. Exploring the choice of k-value effect model performance and fairness.

we conducted experiments using a range of values for $k = \{4, 7, 10, 15, 20, 30, 50\}$, maintaining all other training parameters constant. The comparative analysis of fairCox's predictive performance and fairness under various settings is depicted in Fig. 4. As k increases, fairCox exhibits enhanced performance on *CRIF@k* and *IGD*, indicating more effective optimization for both individual and group fairness. However, the model's predictive accuracy remains largely unaltered when k is modest (*e.g.,* less than 15 for ROSSI, 20 for COMPAS, 15 for KKBox, and 10 for Support), indicating an optimal balance between fairness and performance within this range. Conversely, larger k values increase the number of samples per comparison, introducing more noise and thus resulting in reduced precision due to diminished weights for accurate labels and ambiguous categorizations. In conclusion, a k value of 10 achieves the best measure of fairness and performance.

Ablation Study. To validate the design of fairCox, we conduct ablation studies to examine the impact of varying loss function parts on the model. Specifically, we devise a fairCox variant by assigning a value of μ to zero. Under this configuration, the individual-group fairness loss ($\mathcal{L}_{\text{igf}}$) bears no impact on the total loss, steering the model to only optimize individual fairness. The results are illustrated in Table 3. We can clearly see that although the individual fairness of fairCox-exhibits minor enhancement compared to fairCox, the individual-group fairness and predictive performance decrease significantly. This is because fairCox- overlooks the disparities in individual fairness constraints across different subgroups, resulting in compromised model performance for deprived groups. Overall, these findings highlight the crucial need to take group disparity into consideration when applying individual fairness constraints.

Table 3. Ablation study results for fairCox and fairCox-.

Dataset	CRIF@10%		IGD%		C-index%		Brier%	
	fairCox-	fairCox	fairCox-	fairCox	fairCox-	fairCox	fairCox-	fairCox
ROSSI	53.29	50.63	33.15	48.94	64.42	63.67	14.73	15.07
COMPAS	80.64	77.65	63.47	74.91	70.14	71.47	13.02	12.83
KKBox	68.67	64.47	50.96	62.85	81.71	82.43	14.87	14.45
Support	74.17	72.17	55.83	70.40	69.28	70.69	14.05	13.83

6 Conclusion

In this paper, we introduce two novel fairness notions, specifically devised for generalized censored contexts, to quantify individual unfairness as well as disparities in individual fairness across subgroups. Alongside this, armed with our proposed fairness notions, we present a unified debiasing algorithm designed to mitigate discrimination in scenarios involving censorship to achieve individual fairness with group awareness. Experimental results on on four real-world datasets explicitly include survival information and with socially sensitive concerns validate the effectiveness of our framework with respect to both prediction performance and fairness. Overall, this work not only establishes a new paradigm for achieving individual fairness with group awareness under uncertainty but also paves the way for future research in ML fairness, guiding it toward more applicable and comprehensive approaches.

Acknowledgement. This work was supported in part by the National Science Foundation (NSF) under Grant No. 2245895.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: There's software used across the country to predict future criminals. ProPublica, And it's biased against blacks (2016)
2. Zhang, W., Weiss, J.C.: Fairness with censorship and group constraints. Knowl. Inf. Syst. **65**(6), 2571–2594 (2023)
3. Turner, K., et al.: Longitudinal patient-reported outcomes and survival among early-stage non-small cell lung cancer patients receiving stereotactic body radiotherapy. Radiother. Oncol. **167**, 116–121 (2022)
4. Zhang, W., Weiss, J.: Fair decision-making under uncertainty. In: 2021 IEEE International ICDM. IEEE (2021)
5. Bechavod, Y., Jung, C., Wu, S.Z.: Metric-free individual fairness in online learning. Adv. Neural Inf. Proc. Syst. **33**, 11214–11225 (2020)
6. Bradburn, M.J., Clark, T.G., Love, S.B., Altman, D.G.: Survival analysis part II: multivariate data analysis-an introduction to concepts and methods. Br. J. Cancer **89**(3), 431–436 (2003)

7. Brier, G.W., Allen, R.A.: Verification of weather forecasts. In: Malone, T.F. (ed.) Compendium of Meteorology, pp. 841–848. American Meteorological Society, Boston, MA (1951). https://doi.org/10.1007/978-1-940033-70-9_68
8. Caton, S., Haas, C.: Fairness in machine learning: a survey. ACM Computi. Surv. (2020)
9. Chambless, L.E., Diao, G.: Estimation of time-dependent area under the roc curve for long-term risk prediction. Stat. Med. **25**(20), 3474–3486 (2006)
10. Chinta, S.V., et al.: Optimization and improvement of fake news detection using voting technique for societal benefit. In: 2023 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1565–1574. IEEE (2023)
11. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
12. Chu, Z., et al.: History, development, and principles of large language models-an introductory survey (2024). arXiv preprint arXiv:2402.06853
13. Chu, Z., Wang, Z., Zhang, W.: Fairness in large language models: A taxonomic survey (2024). arXiv preprint arXiv:2404.01349
14. Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G.: Survival analysis part I: basic concepts and first analyses. Br. J. Cancer **89**(2), 232–238 (2003)
15. Cox, D.R.: Regression models and life-tables. J. Roy. Stat. Soc. Ser. B (Methodol.) **34**(2), 187–202 (1972)
16. Diana, E., Gill, W., Kearns, M., Kenthapadi, K., Roth, A.: Minimax group fairness: algorithms and experiments. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 66–76 (2021)
17. Dong, Y., Kang, J., Tong, H., Li, J.: Individual fairness for graph neural networks: a ranking based approach. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 300–310 (2021)
18. Du, M., Liu, N., Yang, F., Hu, X.: Learning credible DNNs via incorporating prior knowledge and model local explanation. Knowl. Inf. Syst. **63**(2), 305–332 (2021)
19. Dwork, C., Hardt, M., Pitassi, T., et al.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226 (2012)
20. Dzuong, J., Wang, Z., Zhang, W.: Uncertain boundaries: Multidisciplinary approaches to copyright issues in generative AI (2024). arXiv preprint arXiv:2404.08221
21. Fleisher, W.: What's fair about individual fairness? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 480–490 (2021)
22. Fox, J., Carvalho, M.S., et al.: The RcmdrPlugin. survival package: Extending the R commander interface to survival analysis. J. Stat. Softw. **49**(7), 1–32 (2012)
23. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
24. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., et al.: Random survival forests. Ann. Appl. Stat. **2**(3), 841–860 (2008)
25. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Med. Res. Methodol. **18**(1), 1–12 (2018)
26. Keya, K.N., Islam, R., Pan, S., Stockwell, I., Foulds, J.: Equitable allocation of healthcare resources with fair survival models. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pp. 190–198. SIAM (2021)
27. Knaus, W.A., Harrell, F.E., Lynn, J., et al.: The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults. Ann. Intern. Med. **122**(3), 191–203 (1995)

28. Kvamme, H., Borgan, Ø., Scheel, I.: Time-to-event prediction with neural networks and cox regression. J. Mach. Learn. Res. **20**(129), 1–30 (2019)
29. Lahoti, P., Gummadi, K.P., Weikum, G.: ifair: Learning individually fair data representations for algorithmic decision making. In: 2019 IEEE 35th international conference on data engineering (icde), pp. 1334–1345. IEEE (2019)
30. Lahoti, P., Gummadi, K.P., Weikum, G.: Operationalizing individual fairness with pairwise fair representations. In: Proceedings of the VLDB Endowment **13**(4) (2019)
31. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisc. Rev. Data Min. Knowl. Discovery **12**(3), e1452 (2022)
32. Long, C., Hsu, H., Alghamdi, W., Calmon, F.: Individual arbitrariness and group fairness. Adv. Neural Inf. Proc. Syst. **36** (2024)
33. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
34. Mukherjee, D., Yurochkin, M., Banerjee, M., Sun, Y.: Two simple ways to learn individual fairness metrics from data. In: International Conference on Machine Learning, pp. 7097–7107. PMLR (2020)
35. Petersen, F., Mukherjee, D., Sun, Y., Yurochkin, M.: Post-processing for individual fairness. Adv. Neural. Inf. Process. Syst. **34**, 25944–25955 (2021)
36. Saxena, N.A., Zhang, W., Shahabi, C.: Missed opportunities in fair AI. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), pp. 961–964. SIAM (2023)
37. Song, W., Dong, Y., Liu, N., Li, J.: Guide: Group equality informed individual fairness in graph neural networks. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1625–1634 (2022)
38. Vasudevan, S., Kenthapadi, K.: Lift: A scalable framework for measuring fairness in ml applications. In: Proceedings of the 29th ACM International Conference on Information &amp; Knowledge Management, pp. 2773–2780 (2020)
39. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1–7. IEEE (2018)
40. Wan, C., Chang, W., Zhao, T., Cao, S., Zhang, C.: Denoising individual bias for fairer binary submatrix detection. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, pp. 2245–2248 (2020)
41. Wang, X., Zhang, W., Jadhav, A., Weiss, J.: Harmonic-mean cox models: a ruler for equal attention to risk. In: Survival Prediction-Algorithms, Challenges and Applications, pp. 171–183. PMLR (2021)
42. Wang, Z., Chu, Z., Blanco, R., Chen, Z., Chen, S.C., Zhang, W.: Advancing graph counterfactual fairness through fair representation learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Nature Switzerland (2024)
43. Wang, Z., Narasimhan, G., Yao, X., Zhang, W.: Mitigating multisource biases in graph neural networks via real counterfactual samples. In: 2023 IEEE International Conference on Data Mining (ICDM), pp. 638–647. IEEE (2023)
44. Wang, Z., Qiu, M., Chen, M., Salem, M.B., Yao, X., Zhang, W.: Towards fair graph neural networks via real counterfactual samples. Knowledge and Information Systems (2024). https://doi.org/10.1007/s10115-024-02161-z
45. Wang, Z., et al.: Preventing discriminatory decision-making in evolving data streams (2023). arXiv preprint arXiv:2302.08017

46. Wang, Z., Wallace, C., Bifet, A., Yao, X., Zhang, W.: Fg^{2}an: fairness-aware graph generative adversarial networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 259–275. Springer Nature Switzerland (2023)
47. Wang, Z., et al.: Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking (2023). arXiv preprint arXiv:2302.08018
48. Yan, S., Kao, H.t., Ferrara, E.: Fair class balancing: enhancing model fairness without observing sensitive attributes. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1715–1724 (2020)
49. Yazdani, S., Saxena, N., Wang, Z., Wu, Y., Zhang, W.: A comprehensive survey of image and video generative AI: recent advances, variants, and applications (2024)
50. Yin, Z., Wang, Z., Zhang, W.: Improving fairness in machine learning software via counterfactual fairness thinking. In: Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings, pp. 420–421 (2024)
51. Zhang, S., et al.: Hidden: hierarchical dense subgraph detection with application to financial fraud detection. In: Proceedings of the 2017 SIAM International Conference on Data Mining, pp. 570–578. SIAM (2017)
52. Zhang, W.: Fairness with censorship: Bridging the gap between fairness research and real-world deployment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 22685–22685 (2024)
53. Zhang, W., Hernandez-Boussard, T., Weiss, J.: Censored fairness through awareness. In: Proceedings of the AAAI conference on artificial intelligence, vol. 37, pp. 14611–14619 (2023)
54. Zhang, W., et al.: Individual fairness under uncertainty. In: 26th European Conference on Artificial Intelligence, pp. 3042–3049 (2023)
55. Zhang, W., Weiss, J.: Longitudinal fairness with censorship. In: Proceedings of the AAAI Conference (2022)
56. Doan, T.V., Chu, Z., Wang, Z., Zhang, W.: Fairness definitions in language models explained (2024)