Redefining Fairness: A Multi-dimensional Perspective and Integrated Evaluation Framework

Zichong Wang[1], Zhipeng Yin[1], Zhen Liu[2],
Roland H. C. Yap[3], Xiaocai Zhang[4], Shu Hu[5], and Wenbin Zhang✉[1]

[1] Florida International University, Miami, USA
[2] Guangdong University of Foreign Studies, Guangzhou, China
[3] National University of Singapore, Singapore
[4] University of Melbourne, Melbourne, Australia
[5] Purdue University, West Lafayette, USA
{ziwang,wenbin.zhang}@fiu.edu

Abstract. As machine learning techniques continue to permeate a variety of application domains with significant societal impact, the focus on algorithmic fairness is becoming an increasingly critical aspect of this established area of research. Existing studies on fairness typically assume that algorithmic bias stems from a single, predefined sensitive attribute in the data, thereby overlooking the reality that multiple sensitive attributes are often prevalent simultaneously in the real world. Unlike previous works, this paper delves into the realm of group fairness involving multiple sensitive attributes, a setting that greatly increases the difficulty of mitigating algorithmic bias. We posit that this multi-attribute perspective provides a more pragmatic model for fairness in real-world applications, and show how learning with such an intricate precondition draws new insights that better explain algorithmic fairness. Furthermore, we develop the first-of-its-kind unified metric, Multi-Fairness Bonded Utility (MFBU), designed to simultaneously evaluate and compare the trade-offs between fairness and utility of multi-source bias mitigation methods. By combining fairness and utility into a single, intuitive metric, MFBU provides model designers the flexibility to holistically evaluate and compare different fairness techniques. Thorough experiments conducted on three real-world datasets substantiate the superior performance of the proposed methodology in minimizing discrimination while maintaining predictive performance.

Keywords: Fairness · Multi-dimensional sensitive attributes · Unified metric · Decision tree.

1 Introduction

Increasing integration of machine learning (ML) algorithms into various information systems for decision-making applications has led to significant successes across numerous domains [14,36,4]. Despite these remarkable achievements, as ML algorithms become more deeply woven into our societal fabric and begin to supplant human decision-making in high-stakes contexts such as resource allocation [68], and loan approval [52], ensuring their fairness has gained increased prominence. This urgency is highlighted

by cases where biases in ML algorithms have resulted in serious consequences, such as Amazon's decision to discard an automated hiring tool biased against women [46] and the predictive policing software PredPol reinforcing racially biased practices by increasing police presence in minority neighborhoods regardless of actual crime rates [27].

In response to these challenges, a number of fairness-aware ML methods have been proposed in recent years that aim to prevent algorithmic decisions from discriminating against specific groups defined by *sensitive attributes* such as race and gender. However, most existing works [68,29,64] addressing fairness in machine learning, including their metrics (*e.g.*, demographic parity [15], equal opportunity [10]), focus primarily on the impact of a single sensitive attribute. This approach ignores that real-world data often contains multiple sensitive attributes such as race and gender operating simultaneously. When fairness is considered along only one dimension, discrimination can persist along others. For example, an algorithm that achieves statistical fairness with respect to race in loan applications might still discriminate against Black female applicants due to unaddressed gender bias, as intersectional combinations of attributes create unique patterns of disadvantage that single-attribute approaches cannot detect [24].

On the other side, in the context of fairness, one long-standing challenge is the trade-off between fairness and predictive performance, where improving fairness typically comes at the cost of reduced accuracy. Current evaluations present fairness improvement and accuracy loss as separate metrics, simply displaying performance indicators and fairness indicators independently. In practical applications, stakeholders need a consolidated metric that includes both dimensions [62]. This challenge becomes even more complex with multiple sensitive attributes, as existing fairness indicators produce multiple values—one for each sensitive attribute—further complicating the evaluation process. Consequently, to address both challenges simultaneously, we need a cohesive metric that can clearly delineate the trade-off between fairness and predictive performance while handling multiple sensitive attributes.

There is an urgent practical need to design decision-making systems and fairness metrics, both of which account for multiple sensitive attributes, presenting unique challenges: **i) Multiple Potential Sources of Bias:** In situations involving multiple sensitive attributes, the bias of a sample may originate from several sources simultaneously [21]. Unlike scenarios with a single sensitive attribute, ensuring fair model predictions in this setting necessitates that all potential sources of bias be mitigated concurrently. **ii) Effective and Robust Trade-off Balance Between Fairness and Performance:** The trade-off between performance and fairness, characterized by the typical inverse relationship between algorithmic fairness based on sensitive attributes and utility [15], introduces additional complexities in multi-sensitive attribute settings. When multiple sensitive attributes are present, each attribute introduces its own fairness-performance trade-off, requiring a balance not just between fairness and performance, but across the different attributes themselves. Consequently, finding an optimal balance between mitigating multiple bias sources and preserving performance becomes a more complex optimization problem that requires a systematic approach. **iii) The Lack of an Intuitive Measure of Model Fairness for Multiple Sensitive Attributes:** Traditional fairness evaluations typically present improvements in fairness and losses in utility as distinct metrics [25]. However, these metrics do not offer a reliable means of jointly quantify-

ing the inherent trade-off between the two. Further, existing methodologies fall short in measuring model fairness under multiple sensitive attributes using a single, unified metric.

To tackle these challenges, *this paper explores the mitigation and quantification of algorithmic bias arising from multiple sensitive attributes, marking the first work, to the best of our knowledge, to simultaneously address both the mitigation and measurement of such biases.* Specifically, we propose a fairness splitting criterion that incorporates bias mitigation for multiple sensitive attributes with an efficient trade-off between utility and fairness. Building on this, we propose a tree-based learning framework to build statistically fair trees for multiple attributes, which can be adapted for any decision tree algorithm. In addition, we propose a unified metric that captures the multifaceted fairness-accuracy trade-off in this complex setting, enabling more direct measurement of fairness and performance trade-offs.

Our major contributions are: **i)** Fair Intersectional Information Gain (FIIG), an innovative splitting criterion designed specifically for fairness-aware ML that pioneers a systematic method to tackle bias across multiple sensitive attributes simultaneously. Our unique splitting criterion seamlessly balances utility and fairness, thereby enhancing both the efficiency and robustness of the model. Furthermore, it can be readily integrated into any decision tree learning algorithm, thus significantly broadening its reach and impact. **ii)** FIIG is further incorporated into a pioneering probabilistic tree learning framework, *Multi-dimensional Fair Decision Tree (MFDT)*, which builds statistically fair trees for multiple sensitive attributes that are flexibly tunable regarding the performance-fairness trade-off. For each node, MFDT generates a Pareto front to first identify the set of Pareto-optimal solutions and then selects the feature maximizing FIIG, thereby extending any decision tree algorithm to balance accuracy and fairness effectively and in an adaptable manner. **iii)** A novel fairness-performance metric, *Multi-Fairness Bonded Utility (MFBU)*, capable of handling multiple sensitive attributes concurrently just as effectively as a single sensitive attribute. MFBU unifies and intuitively evaluates the trade-off between fairness and accuracy when mitigating biases from multiple sources. **iv)** Extensive empirical experiments on three real-world datasets with multiple sensitive attributes demonstrate the efficacy of the proposed unified metrics and fairness-aware algorithm.

2 Related work

Fairness-aware Learning with Single Binary Sensitive Attributes. Fairness is a widespread issue in machine learning systems [25,39]. In recent years, researchers have proposed a number of fairness notions and methods for quantification and mitigation of bias in machine learning algorithms. For instance, Kamiran *et al.* established several key approaches: modifying training data through label and attribute adjustments [21] and developing fairness-aware splitting criteria for decision trees [22]. Building upon these preprocessing and in-processing techniques, Zafar *et al.* [63] advanced the field by incorporating fairness directly into the optimization function, limiting outcome differences across demographic groups while maintaining model integrity. More recent approaches have focused on optimizing fairness more efficiently. MiniMax [28] at-

tempts to reduce maximum group risk, though residual bias can remain, while Herrear *et al.* [35] proposed a novel meta-learning approach using regression models to predict the fairness of hyperparameter settings before full training, reducing computational costs of fairness optimization. Despite these advancements, like most existing AI fairness approaches, these methods are designed for a single binary sensitive attribute (*e.g.*, Gender = {Male, Female}) and struggle to handle scenarios involving multiple sensitive attributes simultaneously.

Fairness-aware Learning with non-Single Binary Sensitive Attributes. Recently, researchers have started exploring fairness beyond binary single sensitive attributes, addressing both multi-valued attributes (*e.g.*, Race = {White, Black, Other}) and multiple attributes simultaneously (*e.g.*, combining Race with Gender). For instance, Morina *et al.* [30] extended the single-attribute fairness metric proposed by [17] and mitigated bias for multiple attributes via a post-processing method. However, their work is limited to binary-sensitive attributes, restricting its applicability in scenarios involving multi-valued attributes like race or ethnicity. Fair-SMOTE [58] is a representative method within another line of approach, data rebalancing techniques, that aim to achieve group fairness by balancing representation among different subgroups in a dataset. This method identifies similarity groups using clustering and generates simulated samples to guarantee adequate representation for all subgroups. However, the oversampling strategy may lead to overfitting due to insufficient numbers of real samples, particularly for intersectional groups with minimal representation. In addition, FairMask [32], a hybrid pre- and post-processing method for multiple attributes, reduces bias from imbalanced training data by using models learned from independent non-sensitive variables to represent sensitive attributes and relabel sensitive attributes seen during deployment. However, they ignore intersectional bias by handling only one sensitive attribute at a time, resulting in samples that may suffer discrimination across different sensitive attributes. For example, Black female applicants may face discrimination based on both gender and race simultaneously.

Different from existing works, our work explicitly addresses intersectional fairness by developing methods that directly handle multiple sensitive attributes simultaneously. Our main contributions are two-fold: First, we propose the Fair Intersectional Information Gain (FIIG) criterion to efficiently tackle bias from multiple sensitive attributes while preserving the advantages of decision trees. This approach offers a unique trade-off between utility and fairness without restricting the algorithm to binary classifiers or specific domains. Second, we introduce the MFBU metric that comprehensively evaluates and compares multiple fairness techniques based on their performance. MFBU facilitates the intuitive selection of fairness techniques for any number of sensitive attributes, accommodating both single-sensitive and multiple-sensitive attribute scenarios. Importantly, our approach is flexible regarding the fairness metrics, enabling end-users to select the most appropriate metric for their specific task.

3 Notations

Given a dataset $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ containing a sequence of independent and identically distributed samples. Each data instance $d_i \in \mathcal{D}$ has an associated class label

$y \in \{0, 1\}$, forming a sequence of class labels Y. The predicted class label is denoted by $\hat{y}$. Every sample d_i can be described as $d_i = \{\mathcal{X}, S, y\}$, where $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ denotes a set of non-sensitive attributes, and $S = \{s_1, s_2, \ldots, s_k\}$ signifies sensitive attributes (*e.g.,* gender, race, age). Specifically, we use $S_0 = \{\forall\ x_i \in \mathcal{X} \mid s_i = 0\}$ to denote the deprived group (*e.g.,* female), and $S_1 = \{\forall\ x_i \in \mathcal{X} \mid s_i = 1\}$ to denote the favored group (*e.g.,* male). Without loss of generality, we consider situations where the given dataset is intrinsically biased with respect to one or more sensitive attributes, which may be binary or multicategorical. Thus, we introduce a set $C = \{c_1, c_2, \ldots, c_a\}$, where each c_i represents a possible combination of sensitive attributes in S.

4 Methodology

This section first introduces our proposed Multi-dimensional Fair Decision Tree, which encompasses: i) a statistical fairness approach for multi-dimensional attributes, ii) multi-dimensional fairness gain that measures bias vary across all pairs of sensitive attribute combinations, and iii) a novel splitting criterion based on a flexible performance-fairness trade-off mechanism. We then present the Multi-Fairness Bonded Utility, which integrates multi-dimensional fairness and performance metrics into a single indicator for comparative analysis, enabling the evaluation of various fairness and performance metric combinations.

4.1 Multi-dimensional Fair Decision Tree

Various fairness-aware approaches have been built upon decision tree models [68,26,20] due to their high interpretability, minimal data preprocessing requirements, computational efficiency, and lack of distributional assumptions; however, like most AI fairness methods, they typically address only single sensitive attributes, overlooking the multi-dimensional nature of bias in real-world scenarios [38]. To this end, we propose the Multi-dimensional Fair Decision Tree (MFDT), a probabilistic tree learning framework designed to: i) **Defining multiple-dimensional Fairness** which accounts for demographic subgroups formed by combinations of multiple sensitive attributes, ii) **Evaluating Multi-dimensional Fairness in Tree Splits** which examines the fairness impact of feature splits across multiple sensitive attribute combinations, enabling us to build decision trees that maintain multi-dimensional fairness throughout the construction process, and iii) **Balance of Multi-dimensional Fairness and Performance** which provides a tunable trade-off between them via Pareto-optimal solutions. The following sections detail each of these components.

i) Defining multi-dimensional Fairness. Straightforward ways to extend existing single-attribute fairness (*e.g.*, statistical parity [15], equal opportunity [19]) to multi-dimensional fairness include summation or maximization of fairness values across individual attributes, but these approaches often overlook intersectional bias. Consider this hypothetical scenario: a bank lending system with age and gender as sensitive attributes. Suppose the system is designed to be unbiased towards gender or age individually. Combining the fairness values of the individual attributes would suggest overall

fairness in the system. However, this fails to capture the complete picture, as multi-dimensional bias towards certain demographic groups can still exist. For instance, the system could favor lending to young men and older women while restricting loans to young women and older men. Examining each sensitive attribute in isolation would miss such multi-dimensional bias. Therefore, it is important to consider these crossover effects when dealing with multi-dimensional fairness [17]. To this end, we extend the statistical fairness notion to quantify model bias under multi-dimensional sensitive attributes, as detailed in Definition 4.1.

Definition 4.1 (Multi-dimensional Fairness). Multi-dimensional Fairness (MF) measures the disparity between different subgroups, where each subgroup is defined by a distinct combination of sensitive attributes (such as "Race & Gender"). We define MF as the maximum statistical disparity in predicted positive outcomes between any two subgroups:

$$MF = \max\left\{\forall c_i, c_j \in C,\, i \neq j : \left|P(\hat{y} = 1 \mid c_i) - P(\hat{y} = 1 \mid c_j)\right|\right\} \tag{1}$$

where c_i represents a distinct combination of sensitive attributes (such as "white female") with C denote the set of all such subgroups.

Overall, multi-dimensional fairness is measured by quantifying the maximum disparity in predicted outcomes between any two subgroups formed by combinations of sensitive attributes. This approach captures intersectional bias that might be ignored when considering sensitive attributes in isolation.

ii) Evaluating Multi-dimensional Fairness in Tree Splits. Although various tree-based fairness splitting criteria have been proposed [68,26,20], they focus solely on a single sensitive attribute, leading to unfair predictions when multiple sensitive attributes are involved. To address this limitation, we first propose *Multi-dimensional Fairness Imparity (MFI)*, which measures fairness disparities across multi-dimensional sensitive attributes simultaneously per the proposed Definition 4.1. Specifically, MFI examines how fairness impacts vary across all pairs of sensitive attribute combinations and identifies the pair with the largest disparity. By doing so, it highlights where a split on a given feature may disproportionately affect certain subgroups, providing a multi-dimensional view of potential disparities. Mathematically, MFI is represented as:

$$\mathrm{MFI}(D, x_j, C) = \max_{\forall c_i,\, c_l \in C,\, i \neq l} \left|\mathrm{MFG}(D, x_j, c_i) - \mathrm{MFG}(D, x_j, c_l)\right| \tag{2}$$

where x_j denotes a feature for splitting dataset D with C representing the set of all subgroups, while MFG refers to Multi-dimensional Fairness Gain, where x_j denotes a feature for splitting dataset D with C representing the set of all subgroups, while MFG refers to Multi-dimensional Fairness Gain, which for each combination of sensitive attributes c_i (*e.g.*, "white female"), is defined as follows:

$$\mathrm{MFG}(D, x_j, c_i) = H(y \mid c_i) - H(y \mid x_j, c_i) \tag{3}$$

where $H(\cdot)$ is the entropy measuring the uncertainty in the distribution of labels, with $H(y \mid c_i)$ represents the entropy of the ground truth labels y given the combination c_i, and $H(y \mid x_j, c_i)$ is the entropy of y after splitting on feature x_j for instances with c_i.

Essentially, a positive value of $MFG(D, x_j, c_i)$ indicates a reduction in uncertainty (*i.e.*, information gain), reflecting potential fairness implications across multi-dimensional demographic subgroups. Although Information Gain (IG) [34] focuses on the overall reduction in uncertainty about Y, MFI extends this concept by highlighting differences in these reductions across various combinations of sensitive attributes. Specifically, MFI measures the gap in uncertainty reductions among different groups defined by c_i that emerge from selecting a feature x_j for splitting. For example, consider two sensitive attribute combinations, such as white males and Black females. MFI would identify whether a split on x_j leads to significantly different information gains for these two groups, revealing potential inequities that would remain hidden when only examining aggregate performance. In this way, MFI goes beyond merely assessing overall utility improvements, capturing changes in predictive performance across different demographic subgroups. While both IG and MFI reward reductions in uncertainty, they differ in perspective: IG prioritizes maximizing accuracy, while MFI concentrates on fairness disparities between all possible combinations of sensitive attributes, highlighting where discrimination may occur.

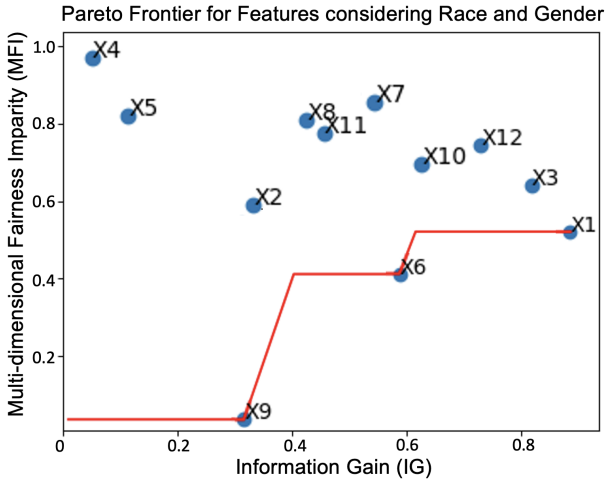

Fig. 1: An illustration of the Pareto Front for balancing utility and fairness.

iii) Balance of Multi-dimensional Fairness and Performance. Another challenge in constructing decision trees is balancing the performance and fairness of each split. To address this, we introduce the concept of the Pareto Frontier [11] into the splitting process of decision trees. In multi-objective optimization scenarios, the Pareto frontier represents the set of all possible optimal solutions that can be obtained without sacrificing any objective. In our model, the objectives are IG and MFI. By incorporating the Pareto frontier, we can better balance performance and fairness during tree construction.

Consider the splitting process for a given node in the decision tree. We have a set of possible splitting attributes for each node that can serve as candidates. Each candidate attribute yields specific IG and MFI values when selected as the split point. Thus, we

can view each candidate attribute as a solution characterized by its IG and MFI scores. Drawing on the concept of the Pareto frontier, we identify all Pareto-optimal solutions. A solution is Pareto-optimal only if no other solution is superior to it on all objectives. In other words, if a solution is considered Pareto-optimal, then we cannot find an improved solution that enhances one objective without deteriorating the score of other objectives. For instance, as Figure 1 shows, attributes X1, X6, and X9 are Pareto-optimal solutions, as X1 achieves better performance than X6 and X9 but has lower fairness. Conversely, X9 achieves better fairness than X1 and X6 but with lower performance. X6 maintains a balance between these two objectives.

After identifying the Pareto-optimal solutions, the Fair Intersectional Information Gain (FIIG) is proposed to select the optimal multi-dimensional fair and accurate splitting as formulated below:

$$\mathrm{FIIG}(\mathrm{D}, \mathrm{X}, \mathrm{C}) = (1 - \alpha) \cdot \mathrm{IG}(\mathrm{D}, \mathrm{X}) - \alpha \cdot \mathrm{MFI}(\mathrm{D}, \mathrm{X}, \mathrm{C}) \tag{4}$$

where $\alpha \in [0, 1]$ is a trade-off parameter to balance the relative importance of utility and fairness in the splitting decision. By optimizing for FIIG rather than just IG, we ensure that the resulting decision tree not only makes accurate predictions but also maintains fairness across intersectional demographic groups. Intuitively, FIIG balances the trade-off between classification performance and fairness: when $\alpha = 0$, FIIG equals IG, prioritizing only classification performance; when $\alpha = 1$, FIIG equals negative MFI, prioritizing only fairness. For values between 0 and 1, FIIG provides a weighted combination of both objectives. This parameter provides flexibility to adjust the model according to specific application requirements, allowing practitioners to appropriately balance utility and fairness based on their domain needs.

Tree Construction. Building upon our multi-objective framework that considers both IG and MFI, we construct the Multi-dimensional Fair Decision Tree (MFDT) by integrating the FIIG into a traditional decision tree workflow. In conventional decision trees (*e.g.*, C4.5 [33]), each node is split by selecting the feature that yields the highest IG. By contrast, we evaluate each candidate feature in our approach using both IG and MFI, generate a Pareto frontier to identify the most balanced solutions, and then apply FIIG (Equation 4) to select the feature that provides the best trade-off between utility and fairness. Specifically, we first compute IG and MFG (and thereby MFI) for all candidate features at a node. We then form the Pareto frontier to filter out any feature dominated by both IG and MFI. From this frontier, we choose the feature that maximizes FIIG, balancing accuracy and fairness through the parameter α. This procedure is repeated at each node until stopping criteria are reached (*e.g.*, purity, feature exhaustion, or minimal node size). Once the splits are determined, the tree is pruned to prevent overfitting, similarly to how pruning is performed in C4.5. However, while conventional pruning only aims to preserve or improve accuracy, our tree structure already incorporates fairness considerations at each node via FIIG. Consequently, even after pruning, MFDT is designed to remain sensitive to disparities across multi-dimensional sensitive attributes.

4.2 Multi-Fairness Bonded Utility

Existing approaches evaluate fairness models by presenting predictive performance and fairness metrics separately through tables, bar charts, or visual comparisons [23,29].

This separation makes it difficult to intuitively assess the trade-off between the two. Moreover, in settings with multi-dimensional sensitive attributes, the fairness metric generates multiple outcomes for distinct sensitive attributes, complicating analysis even further.

To address these challenges, we propose the Multi-dimensional Fairness Bonded Utility (MFBU), which enables the simultaneous evaluation of model performance and fairness through a single consolidated result. Specifically, the MFBU framework consists of three conceptual components that address fundamental challenges in fairness evaluation: **i) Creating a trade-off baseline:** To properly evaluate fairness techniques, we need a standard reference point that reflects inherent trade-offs between performance and fairness. This baseline serves as the foundation for all comparative analyses. **ii) Five effectiveness levels:** Complex numerical metrics alone are difficult to interpret. By categorizing techniques into meaningful effectiveness levels, we enable practitioners to quickly understand the qualitative impact of different approaches without requiring deep statistical knowledge. **iii) Quantitative evaluation of trade-offs:** Beyond categories, precise measurement of trade-offs is necessary for rigorous scientific comparison and optimization. This component allows researchers to quantify the difference between each method. Together, these three components form a comprehensive evaluation framework that bridges the gap between theoretical fairness metrics and practical decision-making. Detailed implementations for each of these components are provided below.

i) Creating a Trade-off Baseline. The foundation of MFBU's trade-off baseline is motivated by the zero-normalization principle proposed by Speicher *et al.* [37], stating that a model's bias is determined by its discriminatory predictions: a model is non-discriminatory if it gives up its predictive power. In other words, a model is not discriminatory if it makes random guesses for each individual, as the predictive performance becomes equally poor across different demographic groups. We use this concept to generate multiple pseudo-models, with a stricter baseline assuming the model makes a single guess that matches the majority label in the dataset. Thus, the model tries to maximize performance while achieving the best fairness. For example, in a loans dataset where 60% of applicants receive a loan and 40% are rejected, the model would be 60% accurate if it predicts everyone will receive a loan.

We visualize this concept by establishing a two-variable coordinate system as shown in Figure 2(a), where the x-axis represents model fairness and the y-axis shows model performance. In this figure, the brown curve tracks how fairness varies across pseudo-models, while the yellow curve depicts corresponding changes in performance. MFBU evaluates models in this two-variable coordinate space, capturing variations across fairness techniques and establishing a trade-off baseline. This approach simplifies decision-making by providing a single consolidated metric that quantifies both performance and fairness, enabling direct comparison across techniques. Within this coordinate system, any combination of metrics can be used: the performance axis can utilize metrics such as Accuracy or F1-score, while the fairness axis can incorporate metrics like Statistical Parity Difference or Equal Opportunity Difference. This flexibility allows MFBU to be tailored to specific application contexts and fairness definitions.

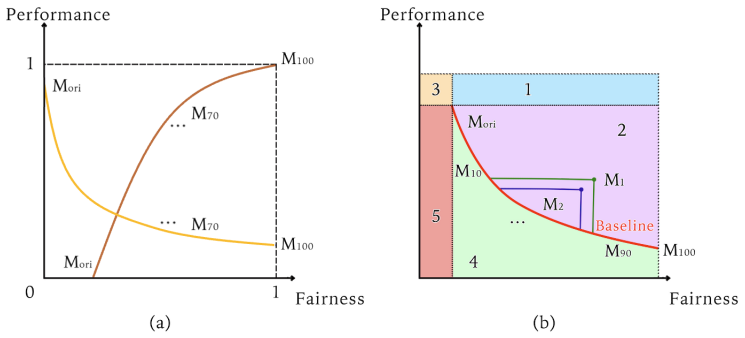

Fig. 2: The MFBU fairness-accuracy trade-off baseline is depicted by the original trade-off point (M_{ori}) and the points generated by the pseudo models (M_{10},. . . , M_{100}). A bias reduction method is considered effective if it shows a superior trade-off compared to the MFBU baseline, *i.e.*, it lies above the red line.

Specifically, MFBU analyze the original model by generating a set of 10 pseudo-models, denoted as M_p, each created by replacing varying percentages (p) of the original model's predictions with consistent output labels to systematically explore fairness improvements. We consider percentages ranging from 10% to 100%. For example, in M_{10}, 10% of the original predictions are randomly selected and replaced with the majority class labels from the input dataset, while in 100%, all predicted labels are replaced with these majority class labels. As illustrated in Figure 2(a), increasing the proportion of replaced predictions leads to improved fairness but simultaneously reduces model accuracy. These pseudo-models provide distinct points along the fairness-performance spectrum, forming the basis for our trade-off baseline analysis. To construct this baseline, we first plot the (performance, fairness) coordinates of the original unadjusted model, labeled as M_{ori} in Figure 2(b). Subsequently, we plot the corresponding coordinates of each pseudo-model (e.g., M_{10}, M_{90}, M_{100}, etc.). By connecting these points, we establish the trade-off baseline, depicted as the red line in Figure 2(b), clearly illustrating the relationship between fairness improvements and performance trade-offs.

For fairness measures designed for a single sensitive attribute, we can directly apply the fairness metric results on the x-axis. However, this is infeasible with multi-dimensional sensitive attributes, as existing fairness metrics generate multiple values for different sensitive attributes. To address this challenge while maintaining a two-dimensional visualization, we need to project these multiple values onto a single axis. Specifically, a model produces only one performance result (such as accuracy or F1-score) regardless of how many sensitive attributes are considered, but generates multiple fairness metrics—one for each sensitive attribute or their combinations—so we project these fairness values for different sensitive attributes onto the space while keeping the performance metric constant. As shown in Figure 3, each dimension represents the fairness value based on a sensitive attribute. We then apply the Euclidean distance vector to calculate a combined fairness result from the various fairness values across multi-dimensional sensitive attributes, effectively representing the multi-dimensional distribution of fairness metrics. This method transforms diverse fairness outcomes into

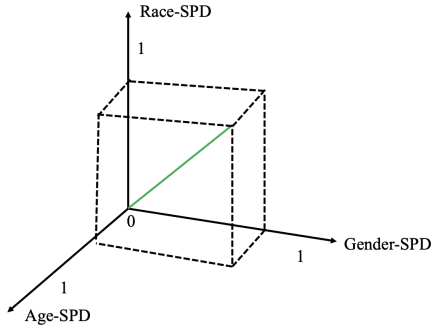

Fig. 3: The MFBU Measure Multi-dimensional Sensitive Attribute

a single indicator, improving the clarity of the trade-off analysis between model performance and fairness. Mathematically, this can be expressed as:

$$\mathrm{F_{multi}} = \sqrt{\omega_1 * {\mathrm{F_1}}^2 + \omega_2 * {\mathrm{F_2}}^2 + \cdots + \omega_\mathrm{n} * {\mathrm{F_n}}^2} \tag{5}$$

where F_i are the single-sensitive-attribute fairness values and ω_i are the corresponding weight parameters with n representing the number of sensitive attributes. These weights adjust the relative importance of different sensitive attributes and can be customized for various application scenarios.

ii) Five Effectiveness Levels. The trade-off baseline provides a framework for categorizing bias mitigation techniques into five distinct effectiveness levels. As shown in Figure 2 (b), Region 1 represents an *Optimal* scenario where a technique improves both model performance and fairness compared to the baseline. Region 2 represents the *Partial Win* scenario where techniques show improved performance or fairness compared to the baseline. Region 3 in Figure 2 (b) represents the *Inverse* scenario, where a technique improves model performance but reduces fairness. The *Partial Loss* scenario is represented by Region 4, where techniques reduce either performance or fairness relative to the baseline. Finally, Region 5 signifies a *Regression* scenario where a technique reduces both performance and fairness compared to the baseline.

iii) Quantitative Evaluation of Trade-offs. The *Optimal*, *Partial Loss*, and *Regression* regions provide clear insights into effectiveness. For a more detailed comparison, we focus on the *Partial Win* category (Region 2). We evaluate different bias mitigation techniques by calculating the area enclosed by the bias-performance points and the baseline. This region, which we call the "Beneficial Balance" region, is shown in Figure 2 (b). Techniques with larger areas are considered better, as they offer more favorable bias-performance balances. We use the area as a metric rather than the distance from the baseline to ensure fair comparison when the baseline curves.

Finally, MFBU produces five percentages for each technique, one for each region, showing the proportion of cases in that region. The total number of cases is calculated as: **Total Cases =** $n_r \times n_t \times n_f \times p_m$, where n_r is the number of run times, n_t is the

number of techniques being compared, n_f is the number of fairness metrics used, and p_m is the number of performance metrics employed.

5 Experiments

Datasets. We conduct experiments on three real-world datasets: i) The **Adult** dataset [18], derived from US census data, is used to predict whether an individual's income exceeds $50K per year based on demographic attributes. Each entry in the dataset corresponds to an individual with information such as education, work class, marital status, and occupation. Race and gender are sensitive attributes. ii) The **COMPAS** dataset [2] is used to predict likelihood of criminal recidivism. Each record corresponds to a criminal defendant with data points such as age, charge degree, and number of priors. The sensitive attributes are the defendant's race and age. iii) The **German credit** dataset [3], used to predict credit risk status, contains credit information from clients of a German bank. Each entry corresponds to an individual with their credit risk categorized as 'good' or 'bad.' The sensitive attributes are age and gender.
Baselines. We compare against four state-of-the-art fairness methods. The first, MiniMax [28], takes a game-theoretic approach to multi-discrimination, formulating it as a mini-max game and aiming for a Pareto efficient solution within a multi-objective problem context. The second, pre-processing method Fair-SMOTE [58], enhances model fairness without requiring direct observation of sensitive attributes. It leverages synthetic minority over-sampling to balance subgroup distribution to improve future predictions' fairness. FairLearn [1], the third baseline, imposes a set of linear fairness constraints on an exponentiated-gradient reduction technique for multi-discrimination. The last baseline, $\mathrm{Kamiran}_{\mathrm{sum}}$ [22], incorporates discrimination awareness directly into the learning process.
Evaluation Metrics. We use accuracy, F1-Score, and the Matthews Correlation Coefficient (MCC)[5] to assess our model. All three can be calculated from a confusion matrix (TP, FP, TN, FN) [25]. Higher accuracy, F1-Score, and MCC values indicate better performance. For fairness evaluation, we measure Statistical Parity Difference (SPD) and Equal Opportunity Difference (EOD), which are widely used metrics [29]. Larger values of SPD and EOD indicate higher levels of bias.

5.1 Experiment Results

RQ1: Does the proposed MFDT help in reducing bias? We compare the performance-fairness trade-off of our proposed method, MFDT, against the five baselines. The results are demonstrated in Table 1. Dark and light blue denote the best and second-best performance, respectively. MFDT achieves superior performance on the Adult dataset in terms of F1-score, and SPD for both race and gender, and EOD for race, showcasing its balanced performance between precision and recall, and lower disparate impact and equality of opportunity difference, thereby reducing bias. It is also successful on the COMPAS dataset, outperforming other models in the F1-score, as well as SPD and EOD for both race and gender. Lastly, on the German credit dataset, MFDT excels in accuracy, MCC, SPD, and EOD for both age and gender, highlighting

Table 1: Performance and fairness comparison of various classification models on real-world datasets - Adult, COMPAS, and Credit. (Dark blue cells denote best and light blue cells denote second-best results.)

Dataset	Methods	Accuracy	F1-Score	MCC	SPD-Race	SPD-Gender	EOD-Race	EOD-Gender
Adult	MiniMax	0.86	0.79	0.56	0.05	0.09	0.06	0.08
	Fair-SMOTE	0.84	0.78	0.57	0.09	0.15	0.05	0.09
	FairLearn	0.83	0.75	0.52	0.08	0.11	0.10	0.09
	$\text{Kamiran}_{\text{sum}} - \text{Race}$	0.78	0.61	0.52	0.17	0.24	0.15	0.19
	$\text{Kamiran}_{\text{sum}} - \text{Gender}$	0.77	0.62	0.54	0.22	0.16	0.17	0.13
	MFDT	0.84	0.79	0.55	0.04	0.08	0.04	0.08
COMPAS	MiniMax	0.82	0.75	0.46	0.10	0.13	0.03	0.02
	Fair-SMOTE	0.66	0.66	0.33	0.12	0.10	0.05	0.02
	FairLearn	0.79	0.71	0.44	0.09	0.10	0.02	0.01
	$\text{Kamiran}_{\text{sum}} - \text{Race}$	0.54	0.48	0.26	0.13	0.16	0.8	0.12
	$\text{Kamiran}_{\text{sum}} - \text{Gender}$	0.51	0.43	0.23	0.13	0.15	0.06	0.11
	MFDT	0.80	0.78	0.45	0.08	0.08	0.02	0.01
German	MiniMax	0.75	0.69	0.41	0.11	0.08	0.03	0.05
	Fair-SMOTE	0.77	0.71	0.43	0.09	0.07	0.05	0.03
	FairLearn	0.75	0.70	0.42	0.05	0.06	0.03	0.04
	$\text{Kamiran}_{\text{sum}} - \text{Age}$	0.73	0.67	0.40	0.15	0.18	0.17	0.22
	$\text{Kamiran}_{\text{sum}} - \text{Gender}$	0.70	0.64	0.41	0.16	0.14	0.18	0.20
	MFDT	0.78	0.70	0.43	0.04	0.06	0.02	0.02

its ability to balance fairness and accuracy across diverse datasets. These results substantiate that our proposed MFDT model effectively reduces bias, as evidenced by its top-ranking performance in terms of SPD and EOD across different protected attributes in diverse datasets. Simultaneously, it maintains competitive accuracy, F1-score, and MCC scores compared to existing methods, indicating a favorable trade-off between fairness and performance. Therefore, we affirm that MFDT is indeed helpful in reducing bias in classification tasks.

RQ2: What is the trade-off of effectiveness between MFDT and other state-of-the-art methods? With the proposed MFBU, we can evaluate the trade-off between Multi-fairness and performance with one illustrative metric. As shown in Figure 4, MFDT consistently outperforms existing methods. MFDT frequently facilitates improvements in both model performance and fairness, as demonstrated with 28% of all cases in the 'Optimal' category, noticeably outperforming other methods. In the 'Partial Win' scenario, which embodies instances where either performance or fairness improved, MFDT accounts for 57% of the cases, outperforming the baseline. It also performs exceptionally well in preventing the 'Regression' trade-off scenario, making up just 1% of cases. Overall, MFDT is a robust and effective approach for managing trade-offs between performance and fairness, especially for multiple sensitive attributes in real-world fairness problems. Further, these findings echo results from RQ1, reaffirming the validity of the MFBU framework. This alignment strengthens our assertion that MFBU is an effective metric for evaluating and comparing performance-fairness trade-offs for multiple bias mitigation methods, proving its effectiveness in real-world applications.

RQ3: How does hyperparameter α impact the balance between classification performance and fairness in MFDT? We answer this based on Figure 5. As α varies between 0 and 1, it significantly changes the model's performance. Specifically, the model's accuracy generally increases as α increases, indicating that the model's perfor-

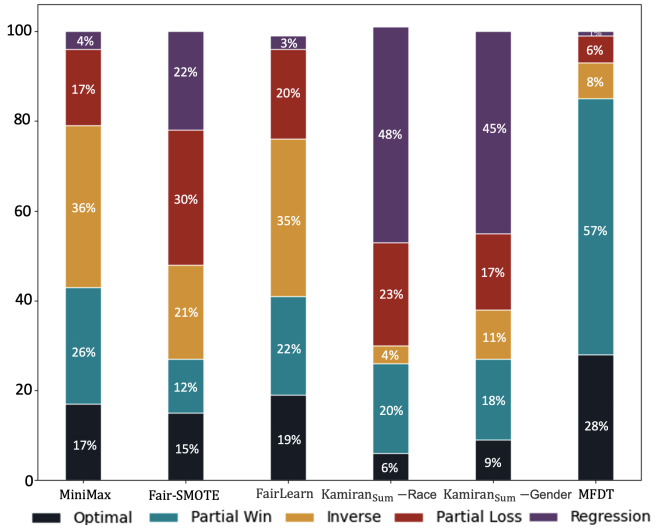

Fig. 4: Different methods' effectiveness distribution in benchmark tasks.

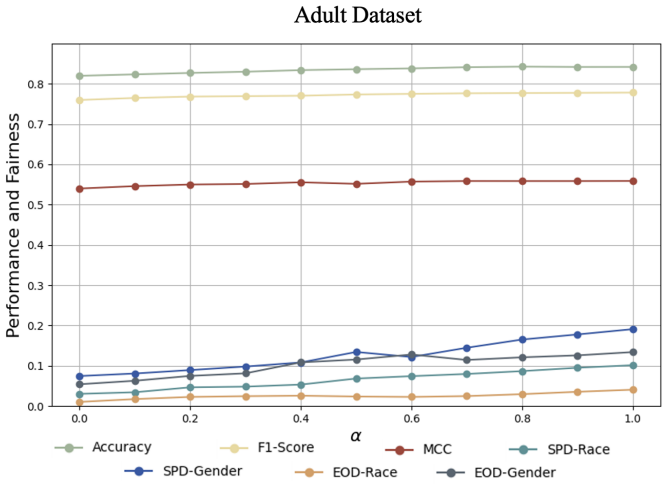

Fig. 5: Effect of α on performance and fairness metrics.

mance improves with larger α. Similarly, the F1 score and MCC exhibit an upward trend with increasing α, implying an enhancement in the model's balance between precision and recall, and its correlation between the observed and predicted binary classifications. The fairness metrics also exhibit an increasing trend with higher α values, indicating a decline in model fairness. However, EOD-Race displays a unique pattern, forming an inverse bell curve, peaking around $\alpha = 0.5$ before declining. This pattern highlights the complex interplay between fairness and accuracy across different α values. In conclusion, there is a trade-off between performance and fairness as α increases. The value of α can be adjusted dynamically to satisfy task-specific requirements.

6 Conclusion

This paper examined multi-dimensional sensitive attributes in fair ML research, a complex challenge that cannot be solved by simply extending single-attribute approaches. We introduced Multi-Fairness Bonded Utility, the first unified metric for evaluating performance-fairness trade-offs among multi-source bias mitigation methods. We proposed Fair Intersectional Information Gain, a novel splitting criterion for fairness-aware decision trees that incorporates Pareto optimality. Our Multi-dimensional Fair Decision Tree provides tunable performance-fairness trade-offs with practical flexibility. Experimental results on real-world datasets validate the effectiveness of our framework with respect to both utility and fairness.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2404039.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning. pp. 60–69. PMLR (2018)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of Data and Analytics, pp. 254–264. Auerbach Publications (2016)
3. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
4. Barata, A.P., Takes, F.W., van den Herik, H.J., Veenman, C.J.: The expose approach to crosslier detection. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 2312–2319. IEEE (2021)
5. Chicco D., J.G.: A statistical comparison between matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index. Journal of Biomedical Informatics p. 104426 (2023)
6. Chinta, S.V., Fernandes, K., Cheng, N., Fernandez, J., Yazdani, S., Yin, Z., Wang, Z., Wang, X., Xu, W., Liu, J., 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)
7. Chinta, S.V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T.L., Zhang, W.: Fairaied: Navigating fairness, bias, and ethics in educational ai applications. arXiv preprint arXiv:2407.18745 (2024)
8. Chinta, S.V., Wang, Z., Zhang, X., Viet, T.D., Kashif, A., Smith, M.A., Zhang, W.: AI-driven healthcare: A survey on ensuring fairness and mitigating bias. arXiv preprint arXiv:2407.19655 (2024)
9. Chu, Z., Wang, Z., Zhang, W.: Fairness in large language models: A taxonomic survey. ACM SIGKDD Explorations Newsletter, 2024 pp. 34–48 (2024)
10. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018)
11. Deb, K.: Multi-objective optimisation using evolutionary algorithms: an introduction. Springer (2011)

12. Doan, T.V., Chu, Z., Wang, Z., Zhang, W.: Fairness definitions in language models explained. arXiv preprint arXiv:2407.18454 (2024)
13. Doan, T.V., Wang, Z., Nguyen, M.N., Zhang, W.: Fairness in large language models in three hours. In: Proceedings of the 33rd ACM International Conference on Information & Knowledge Management (2024)
14. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science advances **4**(1), eaao5580 (2018)
15. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
16. Dzuong, J., Wang, Z., Zhang, W.: Uncertain boundaries: Multidisciplinary approaches to copyright issues in generative AI. arXiv preprint arXiv:2404.08221 (2024)
17. Foulds, J.R., Islam, R., Keya, K.N., Pan, S.: An intersectional definition of fairness. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE). pp. 1918–1921. IEEE (2020)
18. Fox, J., Carvalho, M.S.: The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis. Journal of Statistical Software **49**, 1–32 (2012)
19. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
20. Jeong, H., Wang, H., Calmon, F.P.: Fairness without imputation: A decision tree approach for fair prediction with missing values. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 9558–9566 (2022)
21. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012)
22. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: 2010 IEEE international conference on data mining. pp. 869–874. IEEE (2010)
23. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23. pp. 35–50. Springer (2012)
24. Huan Tian, Bo Liu, Tianqing Zhu, Wanlei Zhou, Philip S. Yu: MultiFair: Model Fairness With Multiple Sensitive Attributes. IEEE Trans. Neural Networks Learn. Syst. 36(3): 5654-5667 (2025)
25. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), e1452 (2022)
26. van der Linden, J., de Weerdt, M., Demirović, E.: Fair and optimal decision trees: A dynamic programming approach. Advances in Neural Information Processing Systems **35**, 38899–38911 (2022)
27. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016)
28. Martinez, N., Bertran, M., Sapiro, G.: Minimax pareto fairness: A multi objective perspective. In: International Conference on Machine Learning. pp. 6755–6764. PMLR (2020)
29. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
30. Morina, G., Oliinyk, V., Waton, J., Marusic, I., Georgatzis, K.: Auditing and achieving intersectional fairness in classification problems. arXiv preprint arXiv:1911.01468 (2019)
31. Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., Beaton, M.: Teens, social media, and privacy. Pew Research Center **21**(1055), 2–86 (2013)
32. J. Ross Quinlan: C4.5: Programs for Machine Learning. Morgan Kaufmann 1993
33. Quinlan, R.: 4.5: Programs for machine learning morgan kaufmann publishers inc. San Francisco, USA (1993)

34. J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.
35. Robles Herrera, S., Monjezi, V., Kreinovich, V., Trivedi, A., Tizpaz-Niari, S.: Predicting fairness of ML software configurations. In: Proceedings of the 20th International Conference on Predictive Models and Data Analytics in Software Engineering. pp. 56–65 (2024)
36. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN computer science **2**(3), 160 (2021)
37. Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2239–2248 (2018)
38. Tan, Y.C., Celis, L.E.: Assessing social and intersectional biases in contextualized word representations. Advances in neural information processing systems **32** (2019)
39. Uddin, S., Lu, H., Rahman, A., Gao, J.: A novel approach for assessing fairness in deployed machine learning algorithms. Scientific Reports **14**(1), 17753 (2024)
40. 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. pp. 40–58. Springer Nature Switzerland (2024)
41. Wang, Z., Chu, Z., Doan, T.V., Wang, S., Wu, Y., Palade, V., Zhang, W.: Fair graph u-net: A fair graph learning framework integrating group and individual awareness. In: proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 28485–28493 (2025)
42. Wang, Z., Dzuong, J., Yuan, X., Chen, Z., Wu, Y., Yao, X., Zhang, W.: Individual fairness with group awareness under uncertainty. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 89–106. 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.: Toward fair graph neural networks via real counterfactual samples. Knowledge and Information Systems pp. 1–25 (2024)
45. Wang, Z., Hoang, N., Zhang, X., Bello, K., Zhang, X., Iyengar, S.S., Zhang, W.: Towards Fair Graph Learning without Demographic Information. The 28th International Conference on Artificial Intelligence and Statistics, vol. 258, pp. 2107–2115 (2025)
46. Wang, Z., Saxena, N., Yu, T., Karki, S., Zetty, T., Haque, I., Zhou, S., Kc, D., Stockwell, I., Wang, X., et al.: Preventing discriminatory decision-making in evolving data streams. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 149–159 (2023)
47. Wang, Z., Ulloa, D., Yu, T., Rangaswami, R., Yap, R., Zhang, W.: Individual fairness with group constraints in graph neural networks. In: 27th European Conference on Artificial Intelligence (2024)
48. Wang, Z., Wallace, C., Bifet, A., Yao, X., Zhang, W.: Fg^2an: 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)
49. Wang, Z., Yin, Z., Zhang, Y., Yang, L., Zhang, T., Pissinou, N., Cai, Y., Hu, S., Li, Y., Zhao, L., et al.: Fg-smote: Towards fair node classification with graph neural network. ACM SIGKDD Explorations Newsletter **26**(2), 99–108 (2025)
50. Wang, Z., Yin, Z., Zhang, Y., Yang, L., Zhang, T., Pissinou, N., Cai, Y., Hu, S., Li, Y., Zhao, L., et al.: Graph fairness via authentic counterfactuals: Tackling structural and causal challenges. ACM SIGKDD Explorations Newsletter **26**(2), 89–98 (2025)
51. Wang, Z., Zhang, W.: Group fairness with individual and censorship constraints. In: 27th European Conference on Artificial Intelligence (2024)

52. Wang, Z., Zhou, Y., Qiu, M., Haque, I., Brown, L., He, Y., Wang, J., Lo, D., Zhang, W.: Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking. arXiv preprint arXiv:2302.08018 (2023)
53. Wang, Z., Wu, A., Moniz, N., Hu, S., Knijnenburg, B., Zhu, Q., Zhang, W.: Towards Fairness with Limited Demographics via Disentangled Learning. In: Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI), (2025)
54. Wang, Z., Zhang, W.: FDGen: A Fairness-Aware Graph Generation Model. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). PMLR, (2025)
55. Wang, Z., Liu, F., Pan, S., Liu, J., Saeed, F., Qiu, M., Zhang, W.: fairGNN-WOD: Fair Graph Learning Without Complete Demographics. In: Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI), (2025)
56. Z., Wang, A., Palikhe, Z., Yin, Z., Zhang, W.: Fairness definitions in language models explained. arXiv preprint arXiv:2407.18454 (2024)
57. Wang, Z., Yin, Z., Yang, L., Zhuang, J., Yu, R., Kong, Q., Zhang, W.: Fairness-Aware Graph Representation Learning Without Demographic Information. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Springer, (2025)
58. 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)
59. 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)
60. Yin, Z., Agarwal, S., Kashif, A., Gonzalez, M., Wang, Z., Liu, S., Liu, Z., Wu, Y., Stockwell, I., Xu, W., et al.: Accessible health screening using body fat estimation by image segmentation. In: 2024 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 405–414. IEEE (2024)
61. Yin, Z., Wang, Z., Xu, W., Zhuang, J., Mozumder, P., Smith, A., Zhang, W.: Digital forensics in the age of large language models. arXiv preprint arXiv:2504.02963 (2025)
62. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: A flexible approach for fair classification. The Journal of Machine Learning Research **20**(1), 2737–2778 (2019)
63. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial intelligence and statistics. pp. 962–970. PMLR (2017)
64. Zhang, W., Bifet, A.: Feat: A fairness-enhancing and concept-adapting decision tree classifier. In: Discovery Science: 23rd International Conference, DS 2020, Thessaloniki, Greece, October 19–21, 2020, Proceedings 23. pp. 175–189. Springer (2020)
65. 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)
66. Zhang, W., Zhou, S., Walsh, T., Weiss, J.C.: Fairness Amidst Non-IID Graph Data: A Literature Review. AI Magazine, vol. 46, no. 1, article e12212 (2025)
67. Zhang, W.: AI Fairness in Practice: Paradigm, Challenges, and Prospects. AI Magazine, vol. 45, no. 3, pp. 386–395 (2024)
68. Zhang, W., Ntoutsi, E.: Faht: an adaptive fairness-aware decision tree classifier. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (2019)
69. Zhang, W., Wang, Z., Kim, J., Cheng, C., Oommen, T., Ravikumar, P., Weiss, J.: Individual fairness under uncertainty. In: 26th European Conference on Artificial Intelligence. pp. 3042–3049 (2023)
70. Zhang, W., Weiss, J.C.: Longitudinal fairness with censorship. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 12235–12243 (2022)