applied sciences

MDPI

Article

Comprehensive Validation on Reweighting Samples for Bias Mitigation via AIF360

Christina Hastings Blow [1,†], Lijun Qian [1], Camille Gibson [2], Pamela Obiomon [1] and Xishuang Dong [1,*,†]

[1] Department of Electrical and Computer Engineering, Texas A&M University System, Prairie View A&M University, Prairie View, TX 77446, USA; chastings@pvamu.edu (C.H.B.); liqian@pvamu.edu (L.Q.); phobiomon@pvamu.edu (P.O.)
[2] Texas Juvenile Crime Prevention Center, College of Juvenile Justice, Texas A&M University System, Prairie View A&M University, Prairie View, TX 77446, USA; cbgibson@pvamu.edu
* Correspondence: xidong@pvamu.edu
† These authors contributed equally to this work.

Abstract: Fairness Artificial Intelligence (AI) aims to identify and mitigate bias throughout the AI development process, spanning data collection, modeling, assessment, and deployment—a critical facet of establishing trustworthy AI systems. Tackling data bias through techniques like reweighting samples proves effective for promoting fairness. This paper undertakes a systematic exploration of reweighting samples for conventional Machine-Learning (ML) models, utilizing five models for binary classification on datasets such as Adult Income and COMPAS, incorporating various protected attributes. In particular, AI Fairness 360 (AIF360) from IBM, a versatile open-source library aimed at identifying and mitigating bias in machine-learning models throughout the entire AI application lifecycle, is employed as the foundation for conducting this systematic exploration. The evaluation of prediction outcomes employs five fairness metrics from AIF360, elucidating the nuanced and model-specific efficacy of reweighting samples in fostering fairness within traditional ML frameworks. Experimental results illustrate that reweighting samples effectively reduces bias in traditional ML methods for classification tasks. For instance, after reweighting samples, the balanced accuracy of Decision Tree (DT) improves to 100%, and its bias, as measured by fairness metrics such as Average Odds Difference (AOD), Equal Opportunity Difference (EOD), and Theil Index (TI), is mitigated to 0. However, reweighting samples does not effectively enhance the fairness performance of K Nearest Neighbor (KNN). This sheds light on the intricate dynamics of bias, underscoring the complexity involved in achieving fairness across different models and scenarios.

Keywords: reweighting samples; bias mitigation; fairness AI; AIF360; traditional machine learning

Citation: Blow, C.H.; Qian, L.; Gibson, C.; Obiomon, P.; Dong, X. Comprehensive Validation on Reweighting Samples for Bias Mitigation via AIF360. *Appl. Sci.* **2024**, *14*, 3826. https://doi.org/10.3390/app14093826

Academic Editors: Milos Manic, Sergiu Dan Stan, Milan Banić, Bogdan Mocan and Ayşegül Uçar

Received: 15 February 2024
Revised: 26 April 2024
Accepted: 27 April 2024
Published: 30 April 2024

1. Introduction

Fairness Artificial Intelligence (AI) endeavors to identify and mitigate bias throughout the entire life cycle of AI technique development, spanning data curation and preparation, modeling, evaluation, and deployment, which is crucial for the successful implementation of trustworthy AI [1]. Bias can manifest in various forms, potentially leading to unfairness in different downstream learning tasks. These biases typically originate from different stages of the Machine-Learning (ML) pipeline, including data curation, algorithm design, and user interactions. Data bias may arise when the data are collected from skewed sources, are incomplete, lack crucial information, or contain errors. Such biases result in unrepresentative or incomplete data, leading to biased outputs. Algorithmic bias stems from biased assumptions or criteria in algorithm design, resulting in biased outputs for downstream tasks. Bias introduced through user interactions occurs when individuals using AI systems inject their own biases or prejudices, whether consciously or unconsciously. To address these sources of bias, various approaches have been proposed. Dataset augmentation involves adding more diverse data to training datasets, enhancing representativeness and

reducing bias [2]. Bias-aware algorithms are designed to consider different types of bias, working to minimize their impact on system outputs [3]. User feedback mechanisms, such as human-in-the-loop systems [4], involve soliciting feedback from users to identify and rectify biases in the system.

Addressing data bias can be efficiently achieved through sample reweighting, contributing to the advancement of fairness in AI [5]. This method involves assigning weights to each sample based on the ratio of its population proportion to its sampling proportion [6]. The process ensures the dataset becomes discrimination-free through two key steps. Firstly, specific attributes, such as gender and race, are identified within the datasets. Subsequently, higher weights are assigned to samples from underrepresented groups, while lower weights are assigned to those from overrepresented groups with respect to these specific attributes. These steps collectively contribute to achieving balance across all groups, thereby fostering fairness in the final outputs of AI algorithms trained on the reweighed data. Nevertheless, existing efforts appear to lack a thorough and comprehensive evaluation of the effectiveness of reweighting samples in mitigating bias associated with traditional machine-learning models.

This paper undertakes a comprehensive evaluation of the effectiveness of sample reweighting in mitigating bias linked with traditional machine-learning models, employing AI Fairness 360 (AIF360) [7]. AIF360 is a versatile open-source library designed to identify and alleviate bias in machine-learning models across the entire AI application lifecycle. It encompasses a comprehensive array of metrics for scrutinizing biases in datasets and models, along with detailed explanations for these metrics and algorithms for bias mitigation. The evaluation in this paper focuses on the reweighting samples methods available in AIF360, applied to classification tasks performed by five traditional machine-learning models: Decision Tree (DT), K Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Random Forest (RF). The reweighting process is carried out with respect to privileged attributes such as sex and race. Subsequently, each traditional machine-learning model undergoes classification tasks on both the original datasets and the new datasets resulting from reweighting samples. The comparative analysis involves assessing the performance of these models on the original and new datasets, respectively. This evaluation is based on metrics such as balanced accuracy and fairness metrics. Experimental results highlight the model-specific nature of reweighting samples effectiveness in achieving fairness in traditional ML models and reveal the complexity of bias dynamics.

The contributions in this paper can be summarized as follows:

1. In contrast to the previous work [8], our research involves a systematic comparison of reweighting samples for mitigating bias on five traditional ML models through the AIF360 platform.
2. We systematically examine the fairness of experimental results with five fairness metrics and provide insights of effectiveness of reweighting samples for bias mitigation.

2. Methodology

This paper aims to examine the effectiveness of reweighting samples to enhance the fairness of traditional machine-learning algorithms on classification tasks. It covers three AI techniques, namely, reweighting samples, traditional machine learning for implementing classification, and fairness metrics to performance evaluation.

2.1. Reweighting Samples

Fairness in AI can be conceptualized as a multi-objective optimization challenge, aiming to optimize learning objectives while mitigating discrimination with respect to sensitive attributes [5]. In essence, achieving fairness may involve a trade-off in learning performance to minimize bias. Data preprocessing emerges as an effective technique for molding training data to foster fairness in AI. Techniques such as suppression, dataset massaging, and reweighting samples have proven effective [5].

Reweighting samples, a specific preprocessing technique, involves adjusting the significance or contribution of individual samples within the training dataset. By strategically assigning weights, it becomes possible to render the training dataset free from discrimination concerning sensitive attributes, all without altering the existing labels. One approach to determining these weights involves measuring them based on the frequency counts associated with the sensitive attribute [9].

This paper leveraged the reweighting samples technique from AIF360 during data preprocessing to enhance fairness in binary classification. The input for the reweighting process comprises a training dataset with samples containing attributes (including a sensitive attribute) and labels, along with the specification of the sensitive attribute. The output is a transformed dataset where sample weights are adjusted concerning the sensitive attribute, mitigating potential classification bias. Throughout the reweighting process, an analysis of the distribution of the sensitive attribute within different groups is conducted. This analysis informs the calculation of reweighting coefficients, which, in turn, adjust sample weights to promote a more uniform distribution across groups.

Given a sensitive (protected) attribute, the privileged group of samples includes the samples with the positive sensitive attribute while the unprivileged group of samples includes the samples with the negative sensitive attribute. For a binary classification task, reweighting samples adjusts the weights of four categories of samples, namely w_{pp} (the weight of the positive privileged sample (pp)), w_{pup} (the weight of positive unprivileged samples (pup)), w_{np} (the weight of negative privileged samples (np)), and w_{nup} (the weight of negative unprivileged samples (nup)), as below.

$$w_{pp} = \frac{N_p}{N_{total}} \times \frac{N_{pos}}{N_{pp}} \tag{1}$$

$$w_{pup} = \frac{N_{up}}{N_{total}} \times \frac{N_{pos}}{N_{pup}} \tag{2}$$

$$w_{np} = \frac{N_p}{N_{total}} \times \frac{N_{neg}}{N_{np}} \tag{3}$$

$$w_{nup} = \frac{N_{up}}{N_{total}} \times \frac{N_{neg}}{N_{up}}, \tag{4}$$

where

N_p: the number of samples in the privileged group.
N_{pp}: the number of samples with the positive class in the privileged group.
N_{np}: the number of samples with the negative class in the privileged group.
N_{up}: the number of samples in the unprivileged group.
N_{pup}: the number of samples with the positive class in the unprivileged group.
N_{nup}: the number of samples with the negative class in the unprivileged group.
N_{pos}: the number of samples with the positive class.
N_{neg}: the number of samples with the negative class.
N_{total}: the number of samples.

2.2. Traditional Machine Learning

This paper applied five traditional machine-learning models to implement binary classification tasks, including LR, DT, KNN, GNB, and RF [10].

LR is a statistical technique employed to model the probability of a binary outcome. It finds widespread application in machine learning for scenarios involving binary classification, aiming to predict whether an instance belongs to one of two classes. The posterior probability of class c_1 is expressed through a logistic sigmoid applied to a linear function of the feature vector ϕ,

$$p(c_1|\phi) = y(\phi) = \sigma(w^T\phi), \tag{5}$$

where $p(c_2|\phi) = 1 - p(c_1|\phi)$ and $\sigma(\cdot)$ is the logistic sigmoid function.

DT is a data mining technique employed to establish classification systems using multiple covariates or to create prediction algorithms for a target variable. This method organizes a population into branch-like segments, forming an inverted tree structure with a root node, internal nodes, and leaf nodes. Notably, the algorithm is non-parametric, allowing it to effectively handle large and complex datasets without imposing a rigid parametric structure [11]. However, one limitation of DT lies in its reliance on hard splits in the input space, where a single model is responsible for predictions for any given value of the input variables.

KNN is a supervised machine-learning algorithm used for classification and regression tasks. It is a type of instance-based learning, where the model makes predictions based on the similarity of new data points to existing labeled data points in the training set [12,13].

GNB is a straightforward classification algorithm [10]. Its primary principle involves assigning labels to classes by maximizing the posterior probability for each sample. This method operates under the assumption that voxel contributions are conditionally independent and follow a Gaussian (normal) distribution. The discriminant function is formulated as the sum of squared distances to the centroid of each class across all voxels in the searchlight. This sum is then weighted by the variance and the logarithm of the a priori probability, computed in the training set using Bayes rule. In essence, GNB provides a probabilistic approach to classification, leveraging assumptions about the distribution of features to make predictions.

RF has proven to be remarkably successful as a classification and regression method. This approach involves the combination of multiple randomized decision trees, and it aggregates their predictions through averaging. Notably, it has demonstrated exceptional performance in scenarios where the number of variables is significantly greater than the number of observations. Its versatility extends to large-scale problems, making it adaptable to various ad hoc learning tasks. Additionally, it provides measures of variable importance, adding to its utility and interpretability [14].

2.3. Fairness Metrics

This paper employed five fairness metrics to evaluate the effectiveness of reweighting samples for mitigating bias.

Given sensitive (protected) attributes, **Disparate Impact (DI)** [7] denotes inadvertent bias that may arise when predictions exhibit varying error rates or outcomes across demographic groups, where the sensitive attributes, such as race, sex, disability, and age, are deemed protected. This bias can emerge either from training models on biased data or from the predictive model itself being discriminatory. In the context of this study, Disparate Impact pertains to divergent impacts on prediction results as defined by

$$DI = \frac{p_{pup}}{p_{pp}}, \tag{6}$$

where p_{pup} refers to the prediction probability for the unprivileged samples with positive predictions while p_{pp} denotes the prediction probability for the privileged samples with positive predictions. If the disparate impact of the predictions approaches 0, it signifies bias in favor of the privileged group. Conversely, if it exceeds 1 it indicates a bias in favor of the unprivileged group. A value of 1 implies perfect fairness in the predictions [15].

Average Odds Difference (AOD) [7] is the average of difference in the False Positive Rates (FPRs) and True Positive Rates (TPRs) between the unprivileged and privileged groups. It is defined by

$$AOD = \frac{(FPR_{up} - FPR_p) + (TPR_{up} - TPR_p)}{2}, \tag{7}$$

where FPR_{up} and FPR_p denote the False Positive Rate for unprivileged and privileged samples, respectively, within predictions, while TPR_{up} and TPR_p refer to the True Positive

Rate for unprivileged and privileged samples, respectively, within predictions. A result of 0 signifies perfect fairness. A positive value indicates bias in favor of the unprivileged group, while a negative value indicates bias in favor of the privileged group.

Statistical Parity Difference (SPD) [7] is used to calculate the difference between the ratio of favorable outcomes in unprivileged and privileged groups. It is defined by

$$SPD = p_{pup} - p_{pp}. \tag{8}$$

A score below 0 suggests benefits for the unprivileged group, while a score above 0 implies benefits for the privileged group. A score of 0 indicates that both groups receive equal benefits.

Equal Opportunity Difference (EOD) [7] involves evaluating the equal opportunity for benefiting all groups. EOD specifically centers on the True Positive Rate (TPR), representing the accurate identification of positives in both the unprivileged and privileged groups. The measure is defined by

$$EOD = TPR_{up} - TPR_{p}. \tag{9}$$

A value of 0 signifies perfect fairness. A positive value indicates bias in favor of the unprivileged group, while a negative value indicates bias in favor of the privileged group.

Theil Index (TI) [7] is also called the entropy index and measures both the group and individual fairness. It is defined by

$$TI = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i}{\mu}ln\frac{b_i}{\mu}, \tag{10}$$

where $b_i = \hat{y_i} - y_i + 1$ and μ is the average of b_i. A lower absolute value of TI value in this context would indicate a more equitable distribution of classification outcomes, while a higher absolute value suggests greater disparity.

3. Results and Discussions

In order to comprehensively validate the reweighting of samples for mitigating classification bias, we utilized five classifiers to conduct classification tasks on two datasets. Subsequently, we assessed the fairness of classification using five evaluation metrics. This enables us to systematically investigate whether reweighting samples effectively mitigates classification bias.

3.1. Dataset

This paper employed two datasets, including the Adult Income dataset and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset, to evaluate the effectiveness of reweighting samples for mitigating fairness.

Adult Income dataset: it includes 48,842 samples with 14 attributes, which can be used for predicting whether income exceeds \$50K/yr based on census data [16].

COMPAS dataset: the COMPAS system is a case management system and decision support tool initially developed and owned by Northpointe (now Equivant). It was designed for the purpose of assessing the likelihood of an individual committing a future crime. The dataset associated with COMPAS comprises more than 20 attributes and includes a substantial sample size of 11,000 instances [17].

3.2. Experimental Metrics

Balance Accuracy (BA) is applicable to both binary and multi-class classification scenarios by computing the mean of sensitivity and specificity. Sensitivity gauges the correct prediction of true positives, representing accurately identified positive instances, while specificity measures the true negatives over the total negatives predicted by the

model. A result nearing 0 signifies poor model performance, whereas a result approaching 1 indicates effective performance across both sensitivity and specificity [18].

3.3. Experimental Results and Discussion

This paper implemented comprehensive validation on reweighting samples for mitigating bias for traditional supervised machine-learning models with the binary classification tasks. It involved two components of experiments on two datasets: the Adult Income dataset and the COMPAS dataset.

3.3.1. Adult Income

Figures 1 and 2 illustrate the performance comparison before and after reweighting samples with respect to the protected attribute *Race*. Prior to reweighting, various ML models exhibit distinct biases in AOD and DI curves. Following reweighting, biases in ML models are mitigated to varying degrees. Notably, DT models demonstrate bias-free behavior in terms of AOD and retain bias in DI values. Conversely, biases in LR and KNN appear unchanged, and GNB produced changed with a high optimal classification threshold, while the bias in RF becomes more unstable. This suggests that the effectiveness of reweighting samples depends on the specific ML model and may not be universally applied.

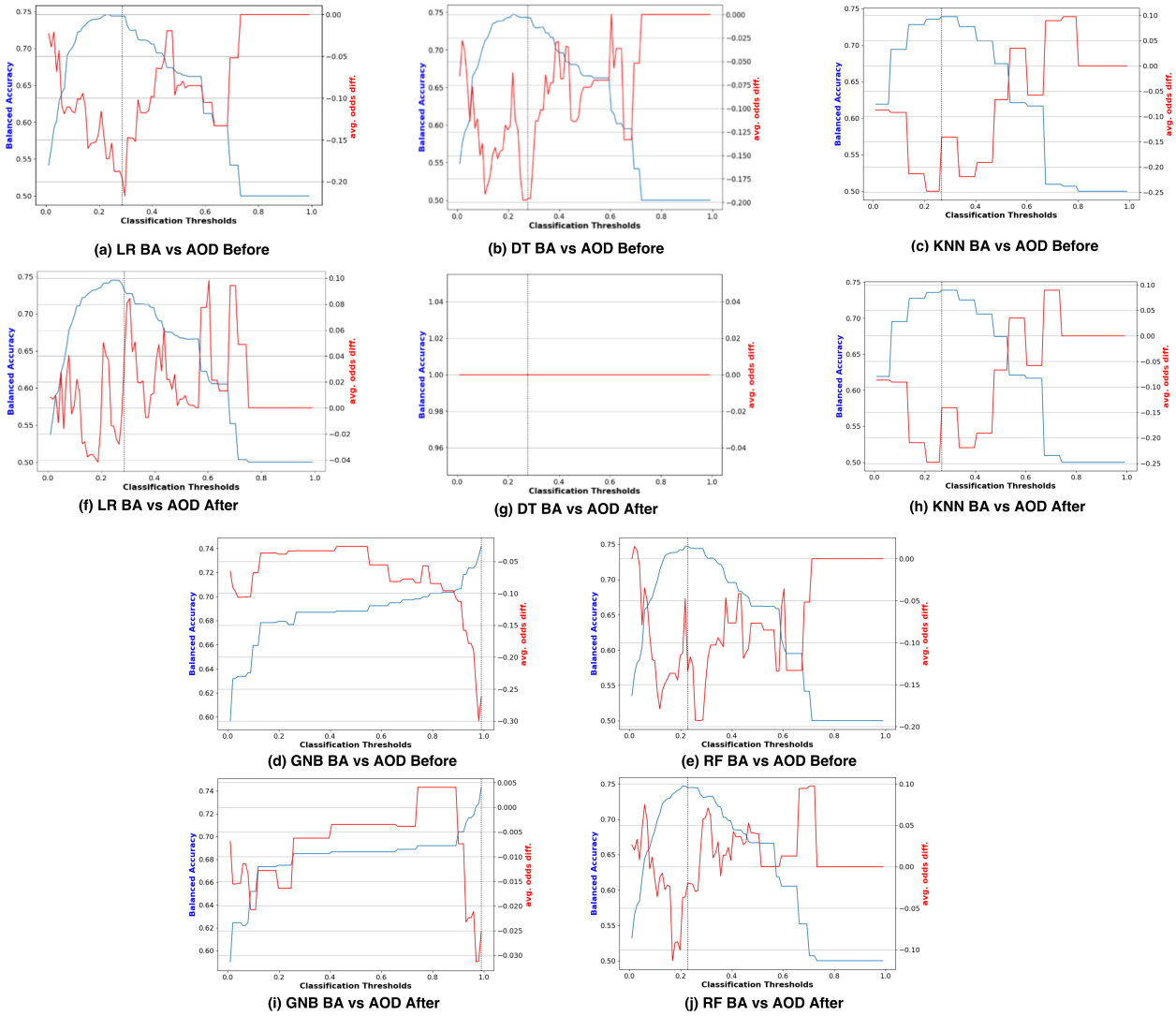

Figure 1. Performance comparison via BA vs. AOD before and after reweighting samples on Adult Income dataset with respect to the protected attribute *Race*.

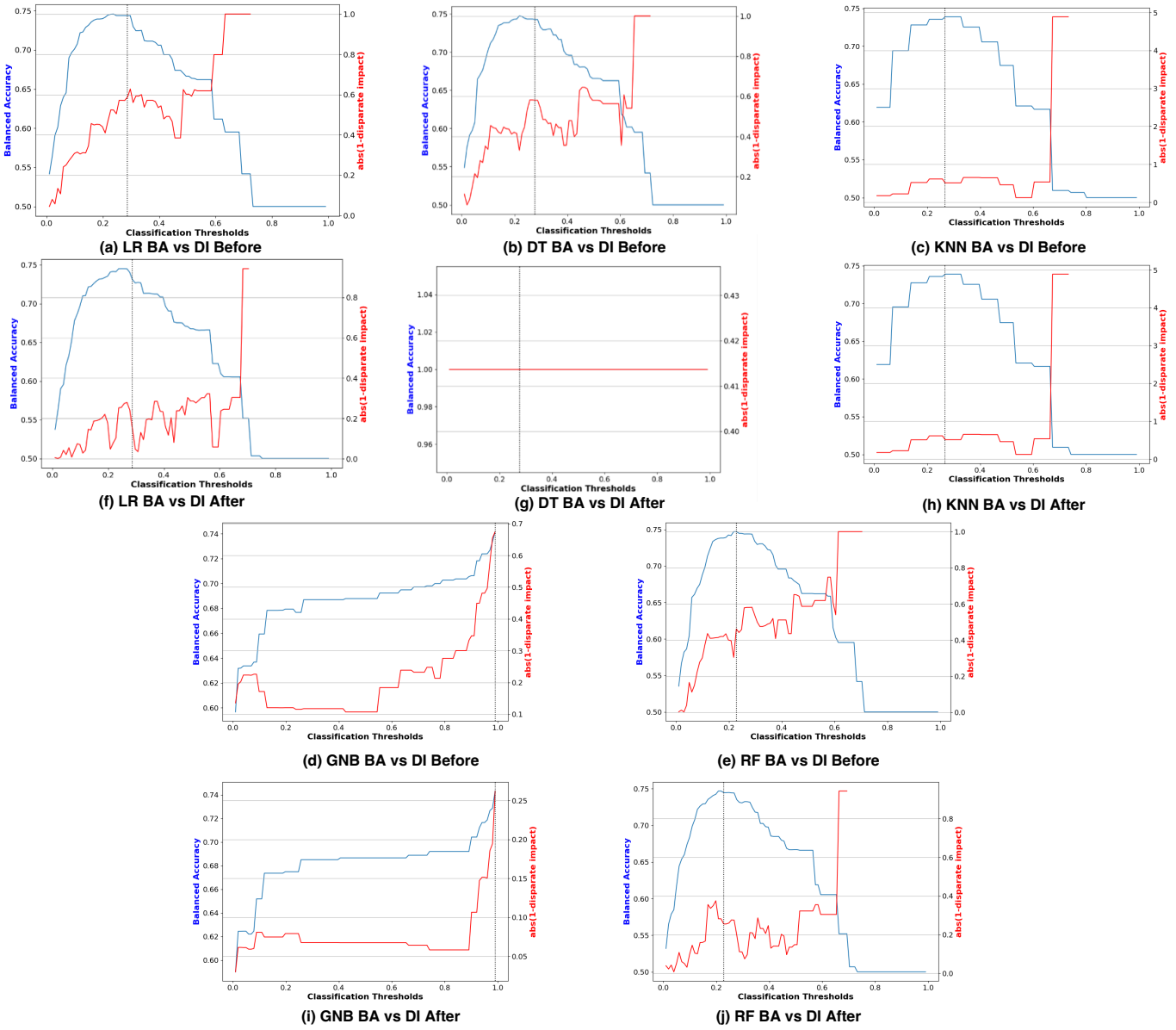

Figure 2. Performance comparison via BA vs. DI before and after reweighting samples on Adult Income dataset with respect to the protected attribute *Race*.

Table 1 provides a systematic comparison using additional fairness metrics. DT's bias is eliminated in terms of AOD, EOD, and TI values. However, regarding SPD and DI values, bias persists towards the unprivileged group, emphasizing the need for a comprehensive examination of bias. Furthermore, although other ML models experience slight reductions in BA, their fairness is marginally improved, emphasizing the inherent trade-off between BA and fairness.

Additionally, Figures 3 and 4 depict comparison results for the protected attribute *Sex*. Similar observations are noted, where reweighting samples are effective mainly for DT, less so for KNN, and have varying impacts on LR, GNB, and RF. Table 2 echoes these trends, revealing slight reductions in BA and modest improvements in fairness for other ML models across various fairness metrics (SPD, AOD, DI, EOD, and TI). This underscores the nuanced effectiveness of the same debiasing technique across different ML models.

Table 1. Performance comparison between before and after reweighting samples through one classification metric, BA, and fairness metrics including SPD, AOD, DI, EOD, and TI on Adult Income dataset regarding the protected attribute *Race*.

Performance before reweighting samples						
Model	BA	SPD	AOD	DI	EOD	TI
DT	0.7426	−0.2416	−0.1959	0.4196	−0.2026	0.1130
GNB	0.7416	−0.2952	−0.2623	0.3252	−0.2872	0.1111
KNN	0.7390	−0.1904	−0.1409	0.4882	−0.1416	0.1207
LR	0.7437	−0.2435	−0.1966	0.4122	−0.2020	0.1129
RF	0.7471	−0.2014	−0.1336	0.5380	−0.1097	0.1066
Performance after reweighting samples						
DT	1.0	−0.1066	0.0	0.5863	0.0	0.0
GNB	0.7432	−0.1147	−0.0252	0.7379	0.0310	0.1058
KNN	0.7390	−0.1904	−0.1409	0.4882	−0.1416	0.1207
LR	0.7311	−0.0523	0.0419	0.8508	0.1083	0.1247
RF	0.7447	−0.1072	−0.0201	0.7449	0.0321	0.1081

Table 2. Performance comparison between before and after reweighting samples through one classification metric, BA, and fairness metrics including SPD, AOD, DI, EOD, and TI on Adult Income dataset regarding the protected attribute *Sex*.

Performance before reweighting samples						
Model	BA	SPD	AOD	DI	EOD	TI
DT	0.7426	−0.3608	−0.3204	0.2785	−0.3775	0.1130
GNB	0.7416	−0.3353	−0.2805	0.3369	−0.3184	0.1111
KNN	0.7390	−0.3983	−0.4075	0.1616	−0.5311	0.1207
LR	0.7437	−0.3580	−0.3181	0.2794	−0.3769	0.1129
RF	0.7471	−0.3777	−0.3292	0.2884	−0.3763	0.1066
Performance after reweighting samples						
DT	1.0	−0.1910	0.0	0.3740	0.0	0.0
GNB	0.7209	−0.0861	0.0073	0.7997	0.0203	0.1192
KNN	0.7390	−0.3983	−0.4075	0.1616	−0.5311	0.1207
LR	0.7134	−0.0705	0.0188	0.7785	0.0293	0.1401
RF	0.7271	−0.1386	−0.0638	0.7220	−0.0774	0.1065

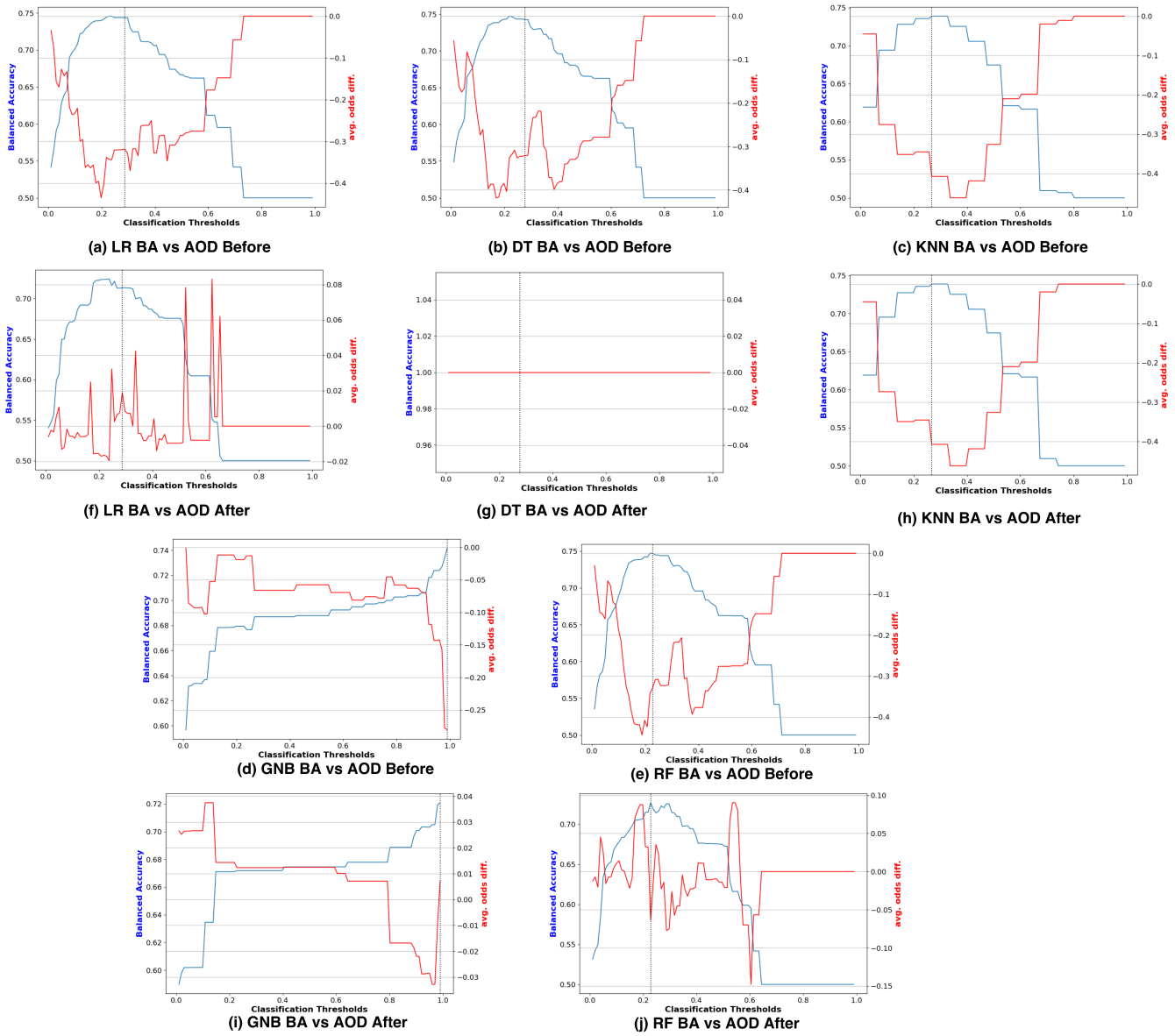

Figure 3. Performance comparison via BA vs. AOD before and after reweighting samples on Adult Income dataset with respect to the protected attribute *Sex*.

3.3.2. COMPAS

To thoroughly assess the effectiveness of reweighting samples in mitigating bias, Figures 5 and 6 provide a performance comparison before and after reweighting samples for the protected attribute *Race* in the COMPAS dataset. Similarly, prior to reweighting, various ML models exhibit diverse biases in the AOD and DI curves. Post-reweighting, DT models demonstrate a lack of bias in AOD while showing significant bias against unprivileged groups in DI values. Conversely, biases in LR, GNB, and RF show significant change over the classification threshold cycle as seen in the curves of AOD and DI. KNN became more bias after reweighing was applied.

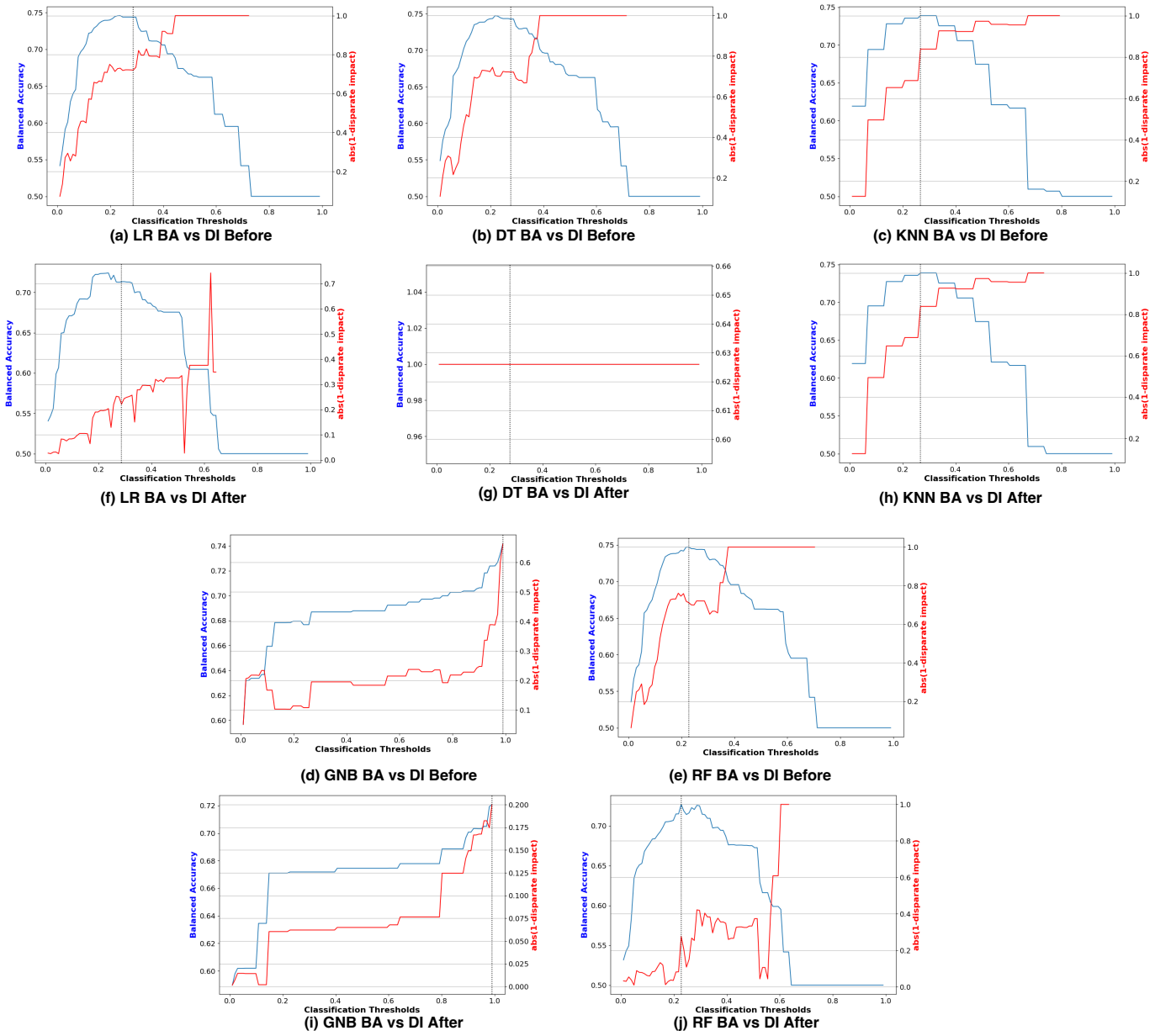

Figure 4. Performance comparison via BA vs. DI before and after reweighting samples on Adult Income dataset with respect to the protected attribute *Sex*.

Table 3 offers a systematic comparison using additional fairness metrics on the COMPAS dataset. Similar to the Adult Income dataset, DT's bias is eradicated concerning AOD, EOD, and TI values. However, biases persist in terms of SPD and DI values, showing a continued bias toward the unprivileged group. The LR and RF models displayed significant bias in favor of the unprivileged group. While GNB and KNN models experience slight reductions in BA, their fairness is marginally improved.

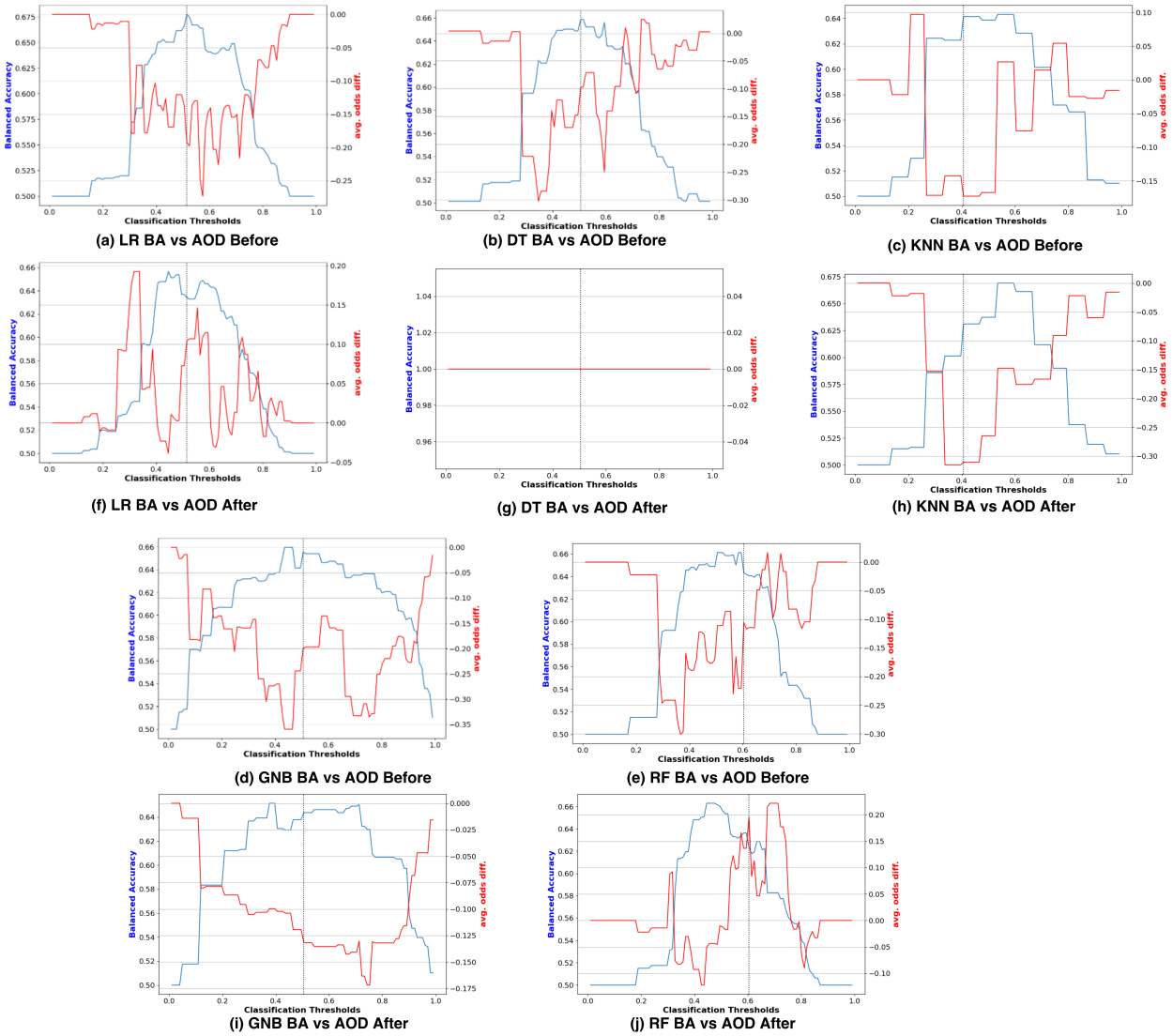

Figure 5. Performance comparison via BA vs. AOD before and after reweighting samples on COMPAS dataset with respect to the protected attribute *Race*.

Table 3. Performance comparison between before and after reweighting samples through one classification metric, BA, and fairness metrics including SPD, AOD, DI, EOD, and TI on COMPAS dataset regarding the protected attribute *Race*.

Performance before reweighting samples						
Model	BA	SPD	AOD	DI	EOD	TI
DT	0.6586	−0.1516	−0.0970	0.7791	−0.1212	0.1835
GNB	0.6553	−0.2483	−0.1994	0.6155	−0.1980	0.2382
KNN	0.6414	−0.2139	−0.1727	0.7282	−0.1131	0.1607
LR	0.6774	−0.2494	−0.1927	0.6600	−0.1877	0.1774
RF	0.6432	−0.1539	−0.1057	0.6873	−0.1166	0.3003
Performance after reweighting samples						
DT	1.0	−0.1769	0.0	0.7060	0.0	0.0
GNN	0.6437	−0.1782	−0.1318	0.6926	−0.1251	0.2594
KNN	0.6311	−0.3432	−0.3105	0.5945	−0.2391	0.1762
LR	0.6342	0.0546	0.1042	1.1062	0.1215	0.2257
RF	0.6234	0.1459	0.1953	1.4012	0.1977	0.2867

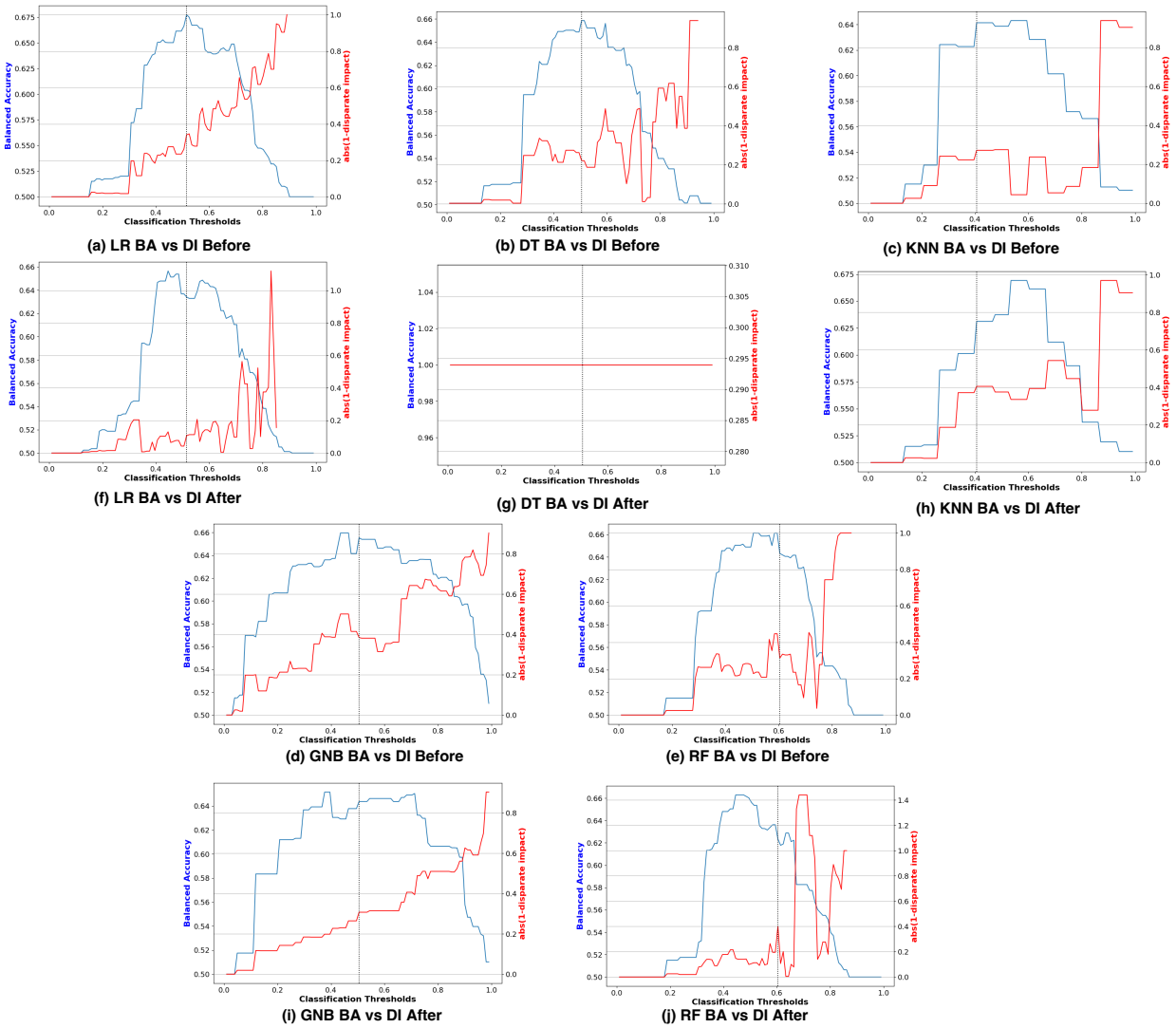

Figure 6. Performance comparison via BA vs. DI before and after reweighting samples on COMPAS dataset with respect to the protected attribute *Race*.

Moreover, Figures 7 and 8 present comparison results for the protected attribute *Sex*. Similar observations are made, where reweighting samples are effective primarily for DT but less so for KNN. Notably, LR, GNB, and RF exhibit more significant effects after reweighting. Table 4 reveals similar trends, with DT showing bias in SPD and DI but less bias overall. All other models bias was reduced while retaining bias against the unpriveleged group. LR showed bias in favor of the unprivileged group for EOD.

In summary, the figures and tables illustrate the impact of reweighting samples on bias mitigation in traditional machine-learning models for both the Adult Income and COMPAS datasets. Notably, DT models showcase effective bias reduction, while KNN exhibits more resistance to debiasing for the Adult Income dataset but work well with the COMPAS dataset. The trade-off between BA and fairness is evident. Results differ across protected attributes, emphasizing the nuanced effectiveness of reweighting. Comprehensive assessments, including additional fairness metrics, reveal the complexity of bias dynamics. This study highlights *the model-specific nature of reweighting sample effectiveness in achieving fairness in traditional machine-learning models*.

Table 4. Performance comparison between before and after reweighting samples through one classification metric, BA, and fairness metrics including SPD, AOD, DI, EOD, and TI on COMPAS dataset regarding the protected attribute *Sex*.

Performance before reweighting samples						
Model	BA	SPD	AOD	DI	EOD	TI
DT	0.6586	−0.1637	−0.1340	0.7759	−0.0597	0.1835
GNB	0.6553	−0.4129	−0.3877	0.5066	−0.3090	0.2382
KNN	0.6414	−0.2336	−0.2095	0.7256	−0.1350	0.1607
LR	0.6774	−0.2724	−0.2439	0.6631	−0.1392	0.1774
RF	0.6432	−0.3759	−0.3484	0.4700	−0.3002	0.3003
Performance after reweighting samples						
DT	1.0	−0.1383	0.0	0.7732	0.0	0.0
GNB	0.6581	−0.1998	−0.1720	0.7154	−0.0899	0.2146
KNN	0.6311	−0.2551	−0.2318	0.7003	−0.1708	0.1762
LR	0.6562	−0.1188	−0.0946	0.8342	0.0111	0.1730
RF	0.6585	−0.1615	−0.1279	0.7081	−0.0760	0.2776

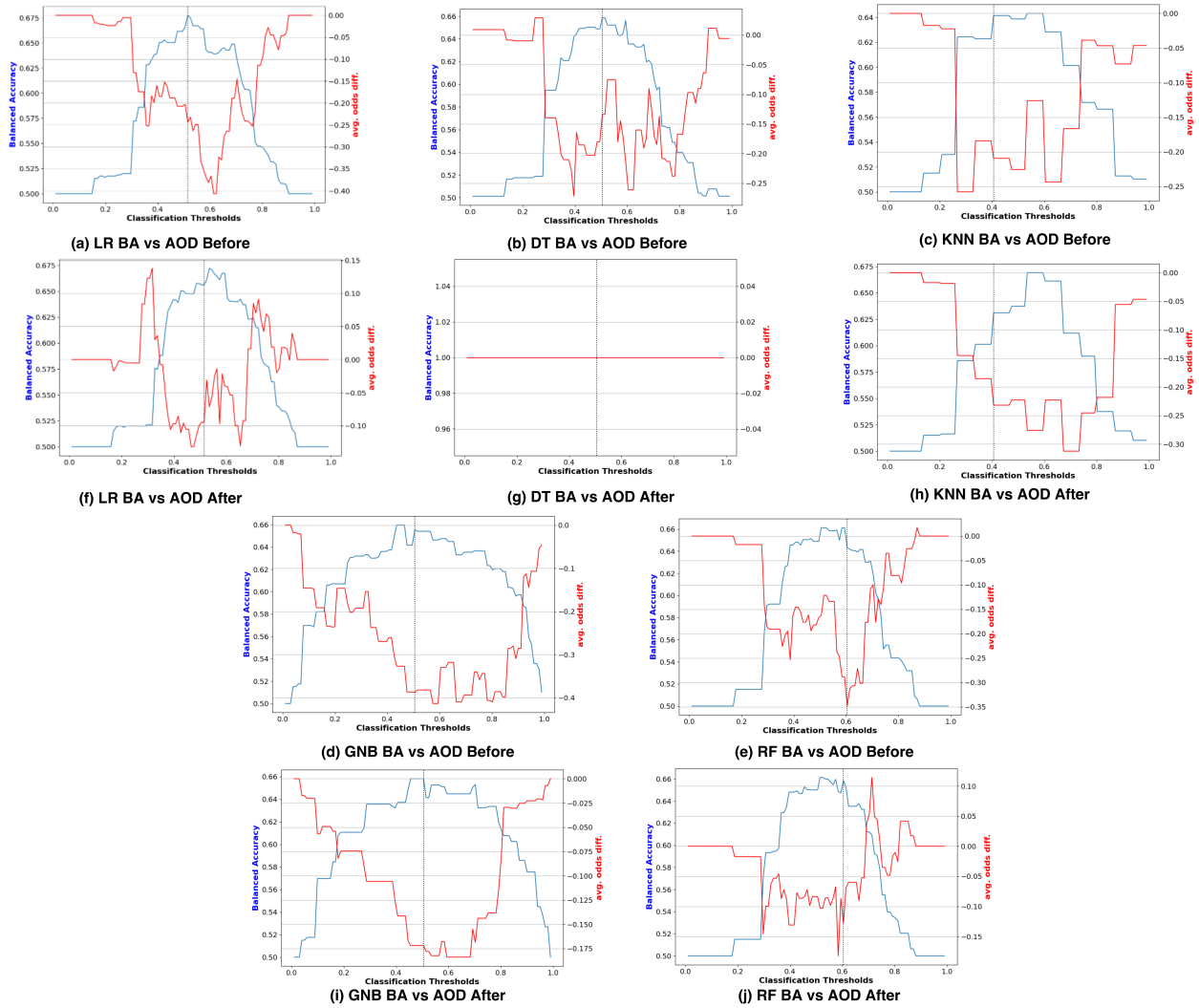

Figure 7. Performance comparison via BA vs. AOD before and after reweighting samples on COMPAS dataset with respect to the protected attribute *Sex*.

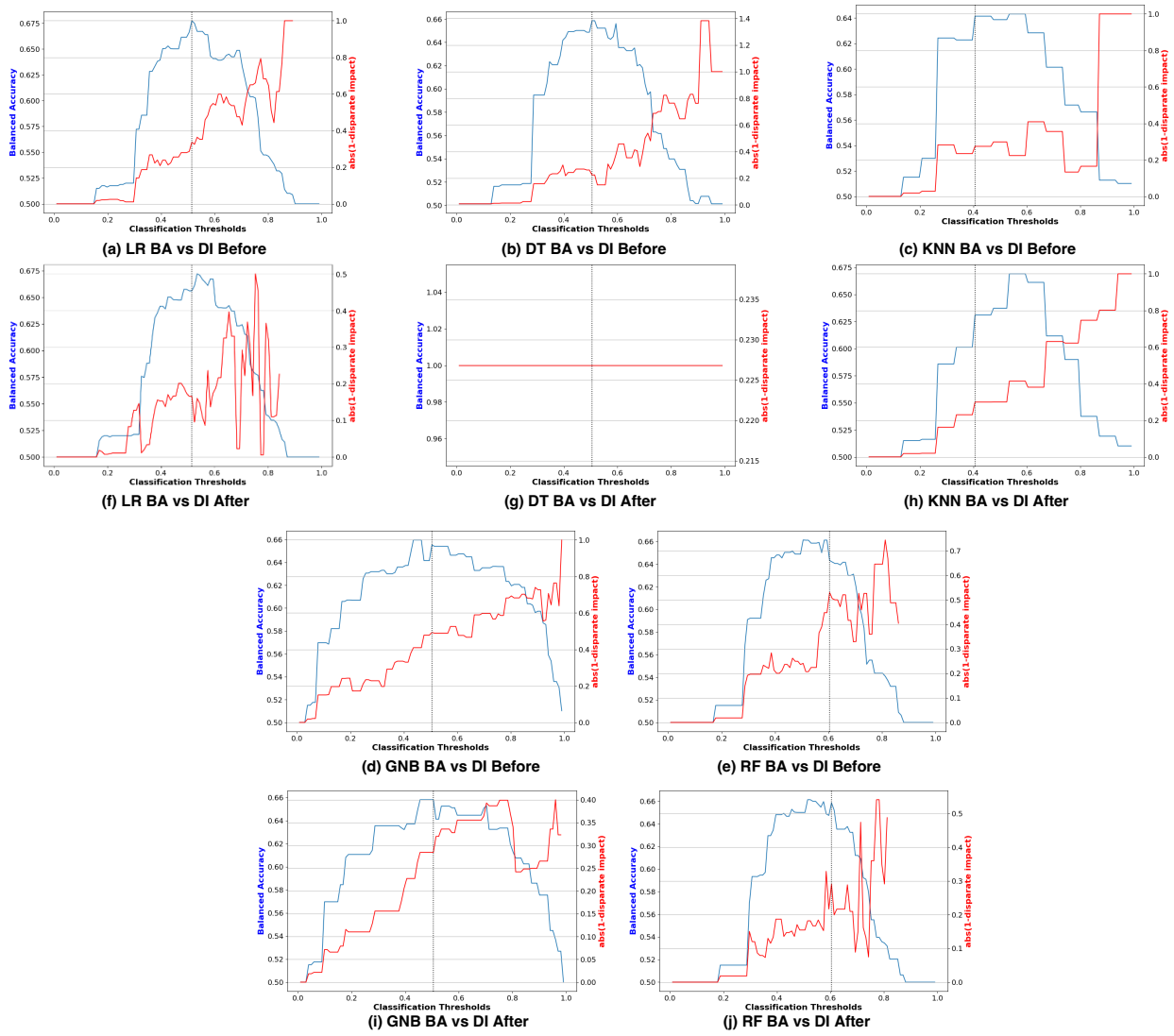

Figure 8. Performance comparison via BA vs. DI before and after reweighting samples on COMPAS dataset with respect to the protected attribute *Sex*.

4. Related Work

ML fairness has become one of the most pivotal challenges of the decade [19]. Intended to intelligently prevent errors and biases in decision-making, ML models sometimes unintentionally become sources of bias and discrimination in society. Concerns have been raised about various forms of unfairness in ML, including racial biases in criminal justice systems, disparities in employment, and biases in loan approval processes [20]. The entire life cycle of an ML model, covering input data, modeling, evaluation, and feedback, is susceptible to both external and inherent biases, potentially resulting in unjust outcomes.

Techniques for mitigating bias in ML models can be categorized into pre-processing, in-processing, and post-processing methods [21]. Pre-processing acknowledges that data itself often introduces bias, with distributions of sensitive or protected variables being discriminatory or imbalanced. This approach modifies sample distributions or transforms data to eliminate discrimination during training [5]. It is considered the most flexible part of the data-science pipeline, making no assumptions about subsequent modeling techniques [22]. In-processing adjusts modeling techniques to counter biases and incorporates fairness metrics into model optimization [23]. Post-processing addresses unfair model outputs, applying transformations to enhance prediction fairness [24].

Pre-processing assumes that the disparate impact of the trained classifier mirrors that of the training data. Techniques include massaging the dataset by adjusting mislabeled class labels due to bias [15,25] and reweighting training samples to assign greater importance to sensitive ones [5,26].

5. Conclusions

Understanding the impact of the biases of machine-learning techniques becomes crucial to prevent unintended behaviors towards specific groups. Fairness AI techniques, such as reweighting samples, have proven effective across various fields This study presents a comprehensive validation of reweighting samples' application to address bias in binary classification using traditional ML models. The comparative analysis involves assessing model performance on original and new datasets, respectively, based on metrics like balanced accuracy and fairness metrics. It offers systematic insights into the effectiveness of different traditional classification algorithms concerning various protected attributes, contributing to the advancement of fairness-enhanced AI. Experimental results underscore the model-specific nature of reweighting sample effectiveness in achieving fairness in traditional ML models, while also revealing the complexity of bias dynamics.

Future research will expand upon this exploration by integrating more advanced machine-learning algorithms, such as generative AI models. For example, generative AI can be utilized for data augmentation with original data to bolster the fairness of AI techniques. Furthermore, the study will integrate new datasets and investigate a wider array of protected attributes to deepen the validation process.

Author Contributions: Conceptualization, P.O., L.Q. and X.D.; methodology, C.H.B., P.O., L.Q. and X.D.; validation, C.H.B., P.O. and X.D.; formal analysis, C.H.B., C.G., P.O., L.Q. and X.D.; data curation, C.H.B. and X.D.; writing—original draft preparation, C.H.B. and X.D.; writing—review and editing, C.H.B., C.G., P.O., L.Q. and X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study involved two datasets, including Adult Income (https://archive.ics.uci.edu/dataset/2/adult, accessed on 10 January 2023) and COMPAS (https://www.kaggle.com/datasets/danofer/compass, accessed on 10 January 2023).

Acknowledgments: This research work is supported by NSF under award number 2323419. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the author(s) and do not necessarily reflect the views of NSF.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIF360	AI Fairness 360
SPD	Statistical parity difference
AOD	Average odds difference
DI	Disparate Impact
EOD	Equal opportunity difference
TI	Theil index
LR	Logistic Regression
DT	Decision Tree
KNN	K Nearest Neighbor
GNB	Gaussian Naive Bayes
RF	Random Forest

References

1. Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From principles to practices. *ACM Comput. Surv.* **2023**, *55*, 1–46. [CrossRef]
2. Sharma, S.; Zhang, Y.; Ríos Aliaga, J.M.; Bouneffouf, D.; Muthusamy, V.; Varshney, K.R. Data augmentation for discrimination prevention and bias disambiguation. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–9 February 2020; pp. 358–364.
3. Ahsen, M.E.; Ayvaci, M.U.S.; Raghunathan, S. When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Inf. Syst. Res.* **2019**, *30*, 97–116. [CrossRef]
4. Ghai, B.; Mueller, K. D-BIAS: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Trans. Vis. Comput. Graph.* **2022**, *29*, 473–482. [CrossRef] [PubMed]
5. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33. [CrossRef]
6. An, J.; Ying, L.; Zhu, Y. Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients. *arXiv* **2020**, arXiv:2009.13447.
7. Bellamy, R.K.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* **2018**, arXiv:1810.01943.
8. Hufthammer, K.T.; Aasheim, T.H.; Ånneland, S.; Brynjulfsen, H.; Slavkovik, M. Bias Mitigation with AIF360: A Comparative Study. 2020. Available online: https://bora.uib.no/bora-xmlui/handle/11250/2764230 (accessed on 10 January 2023).
9. Calders, T.; Kamiran, F.; Pechenizkiy, M. Building classifiers with independency constraints. In Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, Miami, FL, USA, 6 December 2009; IEEE: Piscataway Township, NJ, USA, 2009; pp. 13–18.
10. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4.
11. Song, Y.Y.; Ying, L. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130. [PubMed]
12. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
13. Kramer, O.; Kramer, O. K-nearest neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors*; Springer: Berlin, Germany, 2013; pp. 13–23.
14. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
15. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 259–268.
16. Adult Income Dataset. UCI Irvine Machine Learning Repository. Available online: https://archive.ics.uci.edu/dataset/2/adult (accessed on 10 January 2023).
17. Larson, J.; Surya Mattu, L.K.; Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. 2016. Available online: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed on 10 January 2023).
18. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The balanced accuracy and its posterior distribution. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; IEEE: Piscataway Township, NJ, USA, 2010; pp. 3121–3124.
19. Shaham, S.; Hajisafi, A.; Quan, M.K.; Nguyen, D.C.; Krishnamachari, B.; Peris, C.; Ghinita, G.; Shahabi, C.; Pathirana, P.N. Holistic Survey of Privacy and Fairness in Machine Learning. *arXiv* **2023**, arXiv:2307.15838.
20. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 254–264.
21. Caton, S.; Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* **2020**, *56*, 1–38. [CrossRef]
22. du Pin Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.N.; Varshney, K.R. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 1106–1119. [CrossRef]
23. Jiang, H.; Nachum, O. Identifying and correcting label bias in machine learning. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Online, 26–28 August 2020; pp. 702–712.
24. Kim, M.P.; Ghorbani, A.; Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 247–254.
25. Kamiran, F.; Calders, T. Classifying without discriminating. In Proceedings of the 2009 2nd International Conference on Computer, Control and Communication, Karachi, Pakistan, 17–18 February 2009; IEEE: Piscataway Township, NJ, USA, 2009; pp. 1–6.
26. Doherty, N.A.; Kartasheva, A.V.; Phillips, R.D. Information effect of entry into credit ratings market: The case of insurers' ratings. *J. Financ. Econ.* **2012**, *106*, 308–330. [CrossRef]