FairBalance: Improving Machine Learning Fairness on Multiple Sensitive Attributes With Data Balancing

Zhe Yu
zxyvse@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Joymallya Chakraborty
jchakra@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Tim Menzies
timm@ieee.org
North Carolina State University
Raleigh, North Carolina, USA

ABSTRACT

This paper aims to improve machine learning fairness on multiple sensitive attributes. Machine learning fairness has attracted increasing attention since machine learning software is increasingly used for high-stakes and high-risk decisions. Most existing solutions for machine learning fairness either target only one sensitive attribute (e.g. sex) at a time, or have magic parameters to tune, or have expensive computational overhead. To overcome these challenges, we propose FairBalance to balance the group distribution of training data across every sensitive attribute before training the machine learning models. Our results show that, under the assumption of unbiased ground truth labels, at low computational overhead, FairBalance can significantly reduce fairness metrics (AOD, EOD, and SPD) on every known sensitive attribute without much, if any damage to the prediction performance. In addition, FairBalanceClass, a variant of FairBalance, can balance the class distribution in the training data. With FairBalanceClass, predictions will no longer favor the majority class, thus achieving a higher F_1 score on the minority class. FairBalance and FairBalanceClass also outperform other state-of-the-art bias mitigation algorithms in terms of prediction performance and fairness metrics.

This research will benefit society by providing a simple yet effective approach to improve fairness of machine learning software on data with multiple sensitive attributes. Our results also validate the hypothesis that, on datasets with unbiased ground truth labels, ethical biases in the learned models largely attribute to the training data having (1) difference in group size and (2) difference in class distribution within each group.

To facilitate reuse, reproduction, and validation of this work, our scripts and data are available at https://anonymous.4open.science/r/FairBalance-1E44 under an open-source Apache license (v2.0).

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Software and its engineering** → **Software creation and management**.

KEYWORDS

machine learning fairness, ethics in software engineering

1 INTRODUCTION

With machine learning and artificial intelligence software increasingly being used to make decisions that affect people's lives, much concern has been raised on the fairness in machine learning. Studies have shown that, sometimes the machine learning software behaves in a biased manner that gives undue advantages to a specific group of people (where those groups are determined by sex, race, etc.). Such bias in the machine learning software can have serious consequences in deciding whether a patient gets released from hospital [20, 26], which loan applications are approved [23], which citizens get bail or sentenced to jail [1], who get admitted/hired by universities/companies [13].

Many research has been done trying to mitigate the ethical bias in the machine learning software. However, most existing solutions for machine learning fairness either target only one sensitive attribute (e.g. sex) at a time. For example, on a dataset with two sensitive attributes sex and race, most existing approaches can learn a fair model on sex or a fair model on race, but cannot learn a model which is unbiased on both sex and race [5, 17, 29]. This hinders the application of bias mitigation algorithms since a fair machine learning model cannot be biased on any sensitive attribute.

Some in-processing bias mitigation algorithms can tackle multiple sensitive attributes at the same time by optimizing for both prediction performance and fairness metrics [21]. However, such in-processing bias mitigation algorithms are usually very expensive. Magic parameters also need to be decided beforehand to trade off between prediction performance and fairness metrics. Another recent work from Chakraborty et al. [7] also works with multiple sensitive attributes. This work utilizes an oversampling technique along with a data selection technique to preprocess the data for learning a fairer model. Unfortunately, this approach has expensive computational overhead as well.

In this paper, we propose a simple yet effective algorithm, FairBalance, to learn a fairer machine learning model on every sensitive attribute. FairBalance is a pre-processing technique which balances the training data across every sensitive group so that:

(1) Training data in each **group** have the same total weight.
(2) Class distributions are the same across all **groups**.

Here, each **group** is a possible combination of different values from each sensitive attribute. The hypothesis behind is:

On datasets with unbiased ground truth labels, ethical biases in the learned models largely attribute to the training data having (1) difference in group size and (2) difference in class distribution within each group.

For example, (1) in a certain dataset if the majority of the data belong to a certain gender, the cost for misclassifying data points of the majority gender would be much higher than that of the minority gender, the machine learning model trained on this dataset will be inclined to underfit on data points of the minority gender; (2) in a certain dataset if the majority of a certain gender would be more successful than the other, the machine learning model trained on this dataset will be inclined to believe these falsehoods.

In addition to improving fairness, FairBalanceClass, a variant of FairBalance, also balances the class distribution. This can be very

useful when we care about the model's prediction performance (e.g. precision, recall, and F_1 score) on a minority class.

To validate our hypothesis and the effectiveness of FairBalance, we explore and answer the following research questions with experiments on three commonly used datasets with multiple sensitive attributes and class labels derived from truth:

RQ1: Can FairBalance mitigate machine learning bias against multiple sensitive attributes with low computational overhead? Tested on three widely-used machine learning fairness datasets each with two sensitive attributes, five commonly applied machine learning classifiers trained on the FairBalance preprocessed training data had reduced bias on every sensitive attribute while maintaining similar prediction performances with $O(n)$ computational overhead.

RQ2: Can FairBalance balance class distributions (FairBalanceClass) as well? With the same experiment setup as RQ1, FairBalanceClass achieved higher F_1 score on the minority class while reducing bias on every sensitive attribute. This suggests that class balancing is effective in FairBalanceClass.

RQ3: How does FairBalance perform compared to the existing state-of-the-art bias mitigation algorithms? With the same experiment setup as RQ1, the performance of FairBalance was compared against other state-of-the-art bias mitigation algorithms. As shown in Table 1 (and discussed in Section 5), FairBalance is the only known algorithm that improves fairness on multiple sensitive attributes *without* damaging prediction performance *and* does not require hyperparameter tuning *and* is compatible with different classifiers *and* has far lower computational overhead than other algorithms.

RQ4: Are (1) difference in group size and (2) difference in class distribution within each group the main reasons for machine learning bias when class labels are unbiased? In Section 4, we compared the bias mitigation performance of (0) no data balancing with (1) only balancing the class distributions within each group and (2) balancing both group sizes and class distributions within each group. With results in Section 5, we confirmed that when class labels are unbiased, (2) difference in class distribution within each group is the most important reason for machine learning bias, and (1) difference in group size also contributes to it. Therefore, our hypothesis is valid and these two are the main reasons causing machine learning bias when class labels are unbiased.

Overall, the **significance** of this paper include:

- On datasets with unbiased class labels and multiple sensitive attributes, we show that the proposed approach FairBalance significantly outperformed other state-of-the-art algorithms in terms of reducing ethical bias while maintaining prediction performance. FairBalance can be considered as a standard procedure for training machine learning software on datasets with unbiased class labels and multiple sensitive attributes, given its advantage in performance, low computational overhead ($O(n)$), easy-to-use, and compatibility to most existing machine learning algorithms.
- **RQ4** validated our hypothesis: on datasets with unbiased ground truth labels, ethical biases in the learned models largely attribute to the training data having (1) difference in group size and (2) difference in class distribution within each group. This provides our software engineering community with a better understanding of the causes of ethical bias in machine learning software.

Following the open science policy of ICSE, to facilitate reuse, reproduction, and validation of this work, our scripts and data are available at https://anonymous.4open.science/r/FairBalance-1E44 under an open-source Apache license (v2.0).

The rest of this paper is structured as follows: Section 2 provides the background and related work of this paper. Section 3 introduces our proposed algorithm FairBalance and its variant FairBalanceClass in details. To explore and answer the four research questions, Section 4 presents the experiment setups while Section 5 shows the experiment results. Followed by discussion of limitations and threats to validity in Section 6 and conclusion in Section 7.

2 BACKGROUND AND RELATED WORK

2.1 Scope of This Work

There are two major reasons causing the learned machine learning model to be biased towards certain sensitive groups:

(1) **Biased labels**. Labels in the training data can sometimes be biased already and the model trained on those labels will inherit their bias. This can usually happen in datasets with human decisions as ground truth labels (e.g. predicting whether an HR will hire an applicant).
(2) **Biased learning process**. Even on labels derived from truth (so that the labels can be 100% fair), machine learning algorithms can sometimes learn a biased model.

Some work, e.g. Chakraborty et al. [8], focuses on removing the biased labels but this direction is not in the scope of this paper. This paper focuses on reducing bias in the learning process. We do not question the fairness of labels in our datasets. As a result, we select datasets with labels derived from truth (e.g. whether a person has income higher than 50K). This paper also assumes that all the sensitive attributes are known.

Machine learning researchers have defined several different metrics for assessing whether a trained machine learning model has ethical bias. Among these metrics, FairBalance aims to reduce the difference in the treatments each individual from different groups received from the learned model. Such difference in treatments represent group fairness and are measured by the following three fairness metrics [2]:

- **Statistical Parity Difference (SPD)**: Difference of probability of being assigned to the positive predicted class (PR) for unprivileged and privileged groups (1).

$$SPD = PR_U - PR_P \tag{1}$$

- **Equal Opportunity Difference (EOD)**: Difference of True Positive Rates (TPR) for unprivileged and privileged groups (2).

$$EOD = TPR_U - TPR_P \tag{2}$$

- **Average Odds Difference (AOD)**: Average of difference in False Positive Rates (FPR) and True Positive Rates (TPR) for unprivileged and privileged groups (3).

$$AOD = [(FPR_U - FPR_P) + (TPR_U - TPR_P)] \times 0.5 \tag{3}$$

Where Table 2 and (4) shows the calculation of PR, TPR, and FPR.

Table 1: Comparisons between the proposed approach FairBalance and other state-of-the-art bias mitigation approaches.

Treatment	Reduce bias on single sensitive attribute	Little damage to predictive performance	Reduce bias on multiple sensitive attributes	No hyper-parameters for trade-offs	Low computational overhead	Compatible with different classifiers
FairBalance	✓	✓	✓	✓	✓	✓
Fair-SMOTE [7]	✓	✓	✓	✓		✓
FERMI [21]	✓	✓	✓			
Reweighing [17]	✓	✓		✓	✓	✓
Adversial Debiasing [29]	✓	✓				
Reject Option Classification [18]	✓	✓		✓		✓

Table 2: Combined Confusion Matrix for Privileged (P) and Unprivileged (U) Groups.

	Privileged		Unprivileged	
	Predicted No	Predicted Yes	Predicted No	Predicted Yes
Actual No	TN_P	FP_P	TN_U	FP_U
Actual Yes	FN_P	TP_P	FN_U	TP_U

$$\begin{aligned} PR &= (TP + FP)/(TP + FP + TN + FN) \\ TPR &= TP/(TP + FN) \\ FPR &= FP/(FP + TN) \end{aligned} \tag{4}$$

2.2 Related Work

Prior work in the same scope can be classified into three types depending on the approach applied to remove ethical bias:

Pre-processing algorithms. In this approach, training data is pre-processed in such a way that discrimination or bias is reduced before training the model. Kamiran and Calders [17] proposed *reweighing* method that generates weights for the training examples in each (group, label) combination differently to achieve fairness. Feldman et al. [14] designed *disparate impact remover* which edits feature values to increase group fairness while preserving rank-ordering within groups. Calmon et al. [4] proposed an *optimized pre-processing* method which learns a probabilistic transformation that edits the labels and features with individual distortion and group fairness. Another pre-processing technique, *learning fair representations*, finds a latent representation which encodes the data well but obfuscates information about sensitive attributes [28].

In-processing algorithms. This approach adjusts the way a machine learning model is trained to reduce the bias. Zhang et al. [29] proposed *Adversarial debiasing* method which learns a classifier to increase accuracy and simultaneously reduce an adversary's ability to determine the sensitive attribute from the predictions. This leads to generation of fair classifier because the predictions cannot carry any group discrimination information that the adversary can exploit. Celis et al. [6] designed a *meta algorithm* to take the fairness metric as part of the input and return a classifier optimized with respect to that fairness metric. Kamishima et al. [19] developed *Prejudice Remover* technique which adds a discrimination-aware regularization term to the learning objective of the classifier. Most recently, Lowy et al. [21] measured fairness violation using exponential Rényi mutual information (ERMI) and designed *FERMI*, an in-processing algorithm, to reduce ERMI and prediction errors with stochastic optimization. Lowy et al. [21] showed that *FERMI* outperforms the other state-of-the-art in-processing bias mitigation algorithms on datasets with multiple sensitive attributes.

Post-processing algorithms. This approach adjusts the prediction threshold after the model is trained to reduce specific fairness metrics. Kamiran et al. [18] proposed *Reject option classification* which gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups within a confidence band around the decision boundary with the highest uncertainty. *Equalized odds post-processing* is a technique which particularly concentrate on the Equal Opportunity Difference(EOD) metric [16, 24].

To demonstrate the effectiveness of our proposed approach, FairBalance will be compared against the following selected state-of-the-art bias mitigation algorithms:

Reweighing [17] focuses on binary classification problem with one sensitive attribute. It assigns different weights to the training data as follows:

$$W(s, c) = \frac{|X(s)| \times |X(c)|}{|X(s, c)| \times |X|},$$

where $s \in \{b, w\}$ is the value of the sensitive attribute and $c \in \{Yes, No\}$ is the value of the binary class label.

Fair-SMOTE [7] first utilizes synthetic minority over-sampling technique (SMOTE) [11] to generate synthetic training data points. The training data is divided into subgroups based on class and the sensitive attributes. If class and the sensitive attributes are both binary, there will be 2*2 = 4 subgroups (Favorable & Privileged, Favorable & Unprivileged, Unfavorable & Privileged, Unfavorable & Unprivileged). Initially, these subgroups are of unequal sizes. **Fair-SMOTE** synthetically generates new data points for all the subgroups except the subgroup having the maximum number of data points. As a result, all subgroups become of equal size (same with the maximum one). Note that, in this way, class balancing is also performed. **Fair-SMOTE** then applies Fair Situation Testing—a logistic regression model is trained and then all the data points are predicted by it. After that, the sensitive attribute value for every data point is flipped (e.g. male to female, white to non-white) and tested again to validate whether model prediction changes or not. If the result changes, that particular data point fails the situation testing and will be removed from the training data.

Adversial Debiasing [29] first trains a *predictor* to accomplish the task of predicting C given X. The output layer of the *predictor* is then used as an input to another network called the *adversary* which attempts to predict the value of the sensitive attribute S. Same to a typical GAN [15], the objective is to optimize the weights in the *predictor* so that (1) the error of predicting C given X is minimized and (2) the best performance for the *adversary* predicting S given the output of the *predictor* is also minimized. A tuneable hyperparameter is utilized to tradeoff between (1) and (2). This algorithm is not compatible with other classifiers since it specifies its model structure as GAN.

Reject Option Classification [18] hypothesized that discriminatory decisions are often made close to the decision boundary because of the decision maker's bias. First, a *base classifier* is trained to predict C given X. Then, a critical region is formed based on the prediction probability of the *base classifier*: $P(c|x) - 0.5 < \theta$, where $0.5 < \theta < 1$. For those data points in this critical region, predictions will be generated according to the values of the sensitive attribute, instead of the prediction probabilities: prediction = Yes if S = b (the deprived group); prediction = No if S = w (the favored group). The parameter θ is optimized with tuneable hyperparameters to tradeoff between prediction performance and fairness.

FERMI [21] balances fairness and accuracy with a tuneable hyperparameter λ by learning the following objective function with stochastic gradient descent:

$$\min_{\theta} E_{X,C,S}\{L(X, C; \theta)\} + \lambda D_R(\hat{C}_\theta(X); S), \tag{5}$$

where $\hat{C}_\theta(X)$ is the output of the learned model, $D_R(\hat{C}_\theta(X); S)$ measures the Exponential Rényi Mutual Information (ERMI), and $E_{X,C,S}\{L(X, C; \theta)\}$ measures the prediction errors. This algorithm is not compatible with other classifiers since it specifies a logistic regression classifier with additional objectives of ERMI.

3 FAIRBALANCE

The intuition behind FairBalance is that, a machine learning model will be most likely biased if the training data violates either of the following two balance criteria:

(1) Training data in each **group** have the same total weight.
(2) Class distributions are the same across all **groups**.

Here, a **group** $g \in G$ is a set of data points sharing the same combination of sensitive attributes. For example, all the data points with *sex=male* and *race=white* form one group.

Given that most machine learning algorithms learn to minimize a loss function of misclassification errors, we can analyze the following two violation examples:

(1) Difference in group sizes:

$$|X(g_1)| < |X(g_2)|.$$

The cost for misclassifying data points in g_1 will be lower than that in g_2 since data points in g_1 appear less frequently. As a result, the learned model will tend to predict more accurately on data points in g_2 than g_1. This problem can be resolved by assigning higher weight on misclassification errors for g_1.

(2) Difference in class distributions within each group:

$$|X(g_1, T)|/|X(g_1, F)| < |X(g_2, T)|/|X(g_2, F)|.$$

The learned model will tend to predict more as negatives (resulting in fewer false positives but more false negatives) in g_1 than in g_2, since positive examples appear less frequently in g_1. This also leads to large SPD, AOD, and EOD. This problem can be resolved by adding higher weight on the positive data points in g_1 and negative data points in g_2.

As shown in (6), FairBalance assign different weights to data from different groups to satisfy the above two criteria. FairBalance is different from **Reweighing** since **Reweighing** does not guarantee the first criterion— training data in each **group** g have the same total weight. To validate the importance of this first criterion, in Section 5 we will also compare FairBalance with **Reweighing-Multiple** in (7), an extended version of **Reweighing** on multiple sensitive attributes.

$$W(g, c) = \frac{|X(c)|}{|X(g, c)|}, \qquad \forall g \in G, c \in C. \tag{6}$$

$$W(g, c) = \frac{|X(g)| \times |X(c)|}{|X(g, c)| \times |X|}, \qquad \forall g \in G, c \in C. \tag{7}$$

In addition to balancing the data for fairness, it is sometimes as important to balance the classes in the training data. This is especially true when we care about the prediction performance on the minority class. As Yan et al. [27] stressed, machine learning bias may increase after class balancing, due to the process being oblivious of the inherent properties of the datasets. To resolve the class imbalance problem, we propose FairBalanceClass, which is a variant of FairBalance. FairBalanceClass balances the training data for both fairness and class balance, as shown in (8).

$$W(g, c) = \frac{1}{|X(g, c)|}, \qquad \forall g \in G, c \in C. \tag{8}$$

Algorithm 1 shows the pseudo code for the proposed FairBalance (class_balance = False) and FairBalanceClass (class_balance = True) algorithms. Figure 1 demonstrates how FairBalance and FairBalanceClass assign weights on an example dataset. Note that, given its simplicity, the computational overhead for FairBalance is low— $O(n)$. Although Figure 1 shows an example binary classification problem with two sensitive attributes, FairBalance can be applied to multi-class classification problems with multiple sensitive attributes according to Algorithm 1. FairBalance is also compatible with different classifiers since it only preprocesses the training data.

4 EXPERIMENTS

In this section, we present the experiment setups for answering the four research questions.

4.1 Datasets

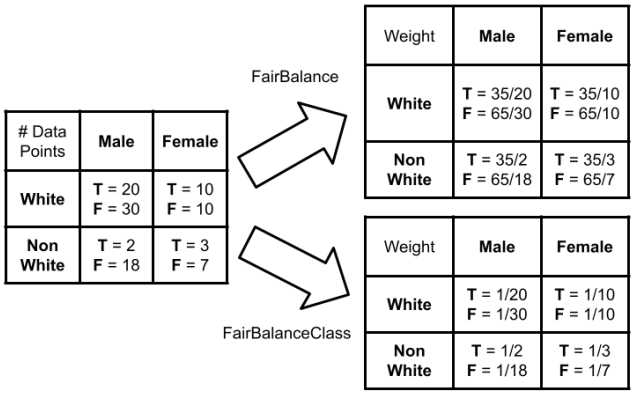
For this study, we selected commonly used datasets in machine learning fairness to conduct our experiments. Starting with datasets seen in recent high-profile papers [7–9, 21], we selected those with multiple sensitive attributes. This leads to the selection of the three datasets shown in Table 3. All three datasets have unbiased ground truth class labels derived from truth, not human decisions, as shown in Table 3.

Following the same pre-processing procedures in Calmon et al. [5] and IBM AIF360 [2], some features, such as "juv_fel_count"

Algorithm 1: FairBalance.

Input : **X**, training data. **G**, groups. **C**, classes. **class_balance**, whether to balance class.
Output : **W(X)**, balanced weights on the training data.

```
if class_balance == True then
    // FairBalanceClass
    for c ∈ C do
        W(c) = 1
else
    // FairBalance
    for c ∈ C do
        W(c) = |X(c)|
for g ∈ G do
    for c ∈ C do
        W(g, c) = W(c) / |X(g) ∩ X(c)|
for x ∈ X do
    // Weights by group and class.
    W(x) = W(x(G), x(C))
return W(X)
```



# Data Points	Male	Female
White	T = 20 F = 30	T = 10 F = 10
Non White	T = 2 F = 18	T = 3 F = 7

Weight	Male	Female
White	T = 35/20 F = 65/30	T = 35/10 F = 65/10
Non White	T = 35/2 F = 65/18	T = 35/3 F = 65/7

Weight	Male	Female
White	T = 1/20 F = 1/30	T = 1/10 F = 1/10
Non White	T = 1/2 F = 1/18	T = 1/3 F = 1/7

Figure 1: Demonstration of FairBalance. In this example, gender and race are the sensitive attributes, T and F are the two possible values of the dependent variable.

in Compas dataset, are dropped while some other features, such as "education-num" in Adult Census Income dataset, are discretized. This makes the number of features different before and after pre-processing as shown in Table 3.

Analysis of the datasets: Figure 2 shows the number of data points in different groups for the three datasets. As we can see, the group sizes are very different in every dataset. Figure 3 clearly shows that for all three datasets, the class distributions within each group are different. Therefore, all three datasets have the two potential problem we identified for machine learning ethical bias in Section 3. We expect that after applying FairBalance, these two problems will be fixed and thus leading to fairer machine learning predictions.

4.2 Experiment Design

We conducted two separate experiments to answer the four research questions. For each experiment, the following process was repeated for 50 times to enable statistical analysis (described in the next subsection).

Our first experiment answers **RQ1** and **RQ2** by comparing FairBalance and its class balancing variant FairBalanceClass against no bias mitigation (**None**). In this experiment, each dataset is randomly split into 70% training data and 30% test data every time. Standard scaler is then applied to normalize both training and test data based on the training data. The three different treatments are applied to the normalized training data:

- **None**: No transformation of the training data.
- **FairBalance**: Assign different weights to data points belonging to different groups to reach the same total weight for each group, as shown in (6) and Algorithm 1.
- **FairBalanceClass**: Assign different weights to data points belonging to different groups and classes to reach the same total weight for each group and class, as shown in (8) and Algorithm 1.

After applying different treatments, five classical classification algorithms are applied to learn from the preprocessed training data and then collect their performances on the test data. The five classification algorithms are:

- **LR:** Logistic Regression Classifier implemented with scikit-learn[1].
- **SVM:** Linear Support Vector Machine Classifier implemented with scikit-learn.
- **DT:** Decision Tree Classifier implemented with scikit-learn.
- **RF:** Random Forest Classifier implemented with scikit-learn.
- **NB:** Gaussian Naive Bayes Classifier from scikit-learn.

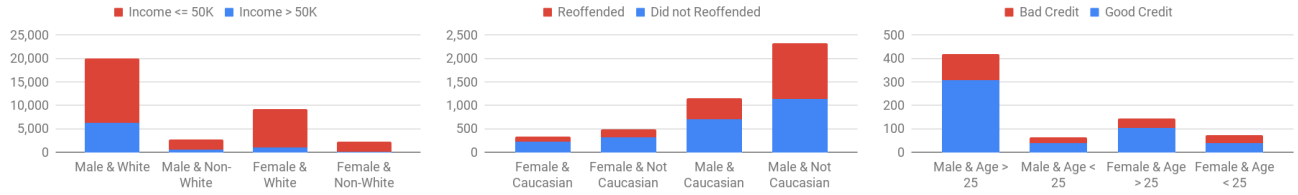
The second experiment answers **RQ3** by comparing FairBalance and FairBalanceClass against five selected state-of-the-art bias mitigation algorithms. In this experiment, each dataset is randomly split into 70% training data and 30% test data every time. Standard scaler is then applied to normalize both training and test data based on the training data. Logistic regression classifier is applied as the base classifier (when the treatment does not specify a classifier). This process was repeated for 10 times for Reject Option Classification due to a memory leak problem in AIF360 [2]. The five baseline bias mitigation algorithms have been introduced in Section 2.2:

- **Reweighing [17]:** The most commonly applied pre-processing bias mitigation algorithm implemented with AIF360[2] under Apache License 2.0.
- **Fair-SMOTE [7]:** The latest open source pre-processing bias mitigation algorithm implemented at https://github.com/joymallyac/Fair-SMOTE under Apache License 2.0.
- **Adversial Debiasing [29]:** An in-processing bias mitigation algorithm implemented with AIF360.
- **Reject Option Classification [18]:** A post-processing bias mitigation algorithm implemented with AIF360.
- **FERMI:** The most state-of-the-art in-processing bias mitigation algorithm which is able to mitigate bias on multiple sensitive attributes simultaneously[3].

[1] https://scikit-learn.org
[2] https://github.com/Trusted-AI/AIF360
[3] https://github.com/optimization-for-data-driven-science/FERMI

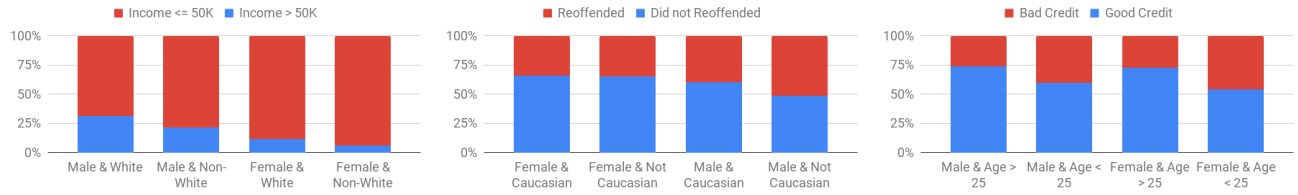
Table 3: Description of the datasets used for the experiment.

Dataset	#Rows	#Features		Sensitive Attributes		Class Labels	
		Before Pre-processing	After Pre-processing	Privileged	Unprivileged	Minority (Target)	Majority
Adult Census Income	48,842	14	18	Sex-Male Race-White	Sex-Female Race-Non-White	Income > 50K 11,687	Income ≤ 50K 37,155
Compas	7,214	28	10	Sex-Female Race-Caucasian	Sex-Male Race-Not Caucasian	Reoffended 2,483	Did not Reoffend 2,795
German Credit	1,000	20	11	Sex-Male Age > 25	Sex-Female Age ≤ 25	Bad Credit 300	Good Credit 700



(a) Adult Census Income **(b) Compas** **(c) German Credit**

Figure 2: Group distribution in absolute values. Group sizes are very different in all three datasets.



(a) Adult Census Income **(b) Compas** **(c) German Credit**

Figure 3: Group distribution in percentage. Class distributions within each group are very different in all three datasets.

Here, Reweighing, Adversial Debiasing, and Reject Option Classification are selected due to their popularity in each category and the fact that they have already been implemented in IBM AIF360 [2]. Fair-SMOTE is selected since it is a similar pre-processing algorithm and is proposed in a recent high-profile paper [7] in software engineering. FERMI [21] is selected since it is the most state-of-the-art in-processing to handle multiple sensitive attributes.

Since most of the above bias mitigation algorithms (except for **FERMI**) are designed for data with single sensitive attribute, we also include the following variants of **Reweighing** and **Fair-SMOTE**:

- **Reweighing-Multiple: Reweighing** only works for single sensitive attribute when proposed [17]. However, it is easy to extend it to multiple sensitive attributes, as described in (7). Compared to FairBalance, this algorithm only balances class distributions within each group, does not balance the group sizes. Therefore, comparing the performance of **Reweighing-Multiple** with **None** will show the effectiveness of balancing class distributions within each group. Comparing the performance of **Reweighing-Multiple** with **FairBalance** will show the effectiveness of balancing the group sizes. **RQ4** can thus be answered with these comparisons.
- **Fair-SMOTE-Multiple:** an extended version of Fair-SMOTE to data with multiple sensitive attributes [7]. This will be the first time **Fair-SMOTE-Multiple** being thoroughly tested on datasets with multiple sensitive attributes.

RQ4 will be answered by comparing **Reweighing-Multiple** (fixing only the second balance criterion) with **FairBalance** (fixing both balance criteria) and **None** (fixing none of the balance criteria).

4.3 Evaluation

The three machine learning fairness metrics (AOD, EOD, and SPD) described in Section 2.1 are applied to evaluate how bias each treatment is on every sensitive attribute. In the meantime, accuracy is applied to evaluate the overall prediction performance and F_1 score is applied to evaluate the prediction performance on the minority class:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (9)$$

$$Precision = TP/(TP + FP)$$

$$Recall = TPR = TP/(TP + FN) \quad (10)$$

$$F_1 = 2 \times Precision \times Recall/(Precision + Recall)$$

Here for the F_1 score, the minority class is treated as the target class "Yes" in the confusion matrix.

Each treatment is evaluated multiple times during experiments as described in the previous subsection. Medians (50th percentile) and IQRs (75th percentile - 25th percentile) are collected for each performance metric since the resulting metrics do not follow a normal distribution. In addition, a nonparametric null-hypothesis significance testing (Mann–Whitney U test [22]) and a nonparametric effect size testing (Cliff's delta [12]) are applied to check if one treatment performs significantly better than another in terms of a specific metric. A set of observations is considered to be significantly different from another set if and only if the null-hypothesis is rejected in the Mann–Whitney U test and the effect size in Cliff's delta is medium or large.

Similar to the Scott-Knott test [25], rankings are also calculated to compare different treatments with nonparametric performance results. For each metric, the treatments are first sorted by their median values in that metric. Then, each pair of treatments are compared with the Mann–Whitney U test and Cliff's delta to decide whether they should belong to the same rank. The pseudo code for the ranking algorithm (assuming lower the better for the metric) is shown in Algorithm 2 in the Appendix.

Algorithm 2: Nonparametric ranking.

Input : **T**, performances to rank, a list of list.
Output : **R**, rankings of the each treatment (row) in **T**.

```
medians = []
for t ∈ T do
    medians.append(median(t))
asc = argsort(medians)
base = T[asc[0]]
rank = 0
R = []
R[asc[0]] = 0
for i=1, i<m, i++ do
    if MannWhitneyU(T[asc[i]], base) < 0.05 &
     CliffsDelta(T[asc[i]], base) > 0.33 then
        rank = rank + 1
        base = T[asc[i]]
    R[asc[i]] = rank
return R
```

5 RESULTS

In this section, we present the experimental results following the setups in Section 4.

RQ1. Can FairBalance mitigate machine learning bias against multiple sensitive attributes with low computational overhead?

Table 4 shows the results of the first experiment. Comparing the performance of FairBalance with None, we observe:

- F_1 score of FairBalance is usually comparable to that of None, except for the german dataset. Accuracy of FairBalance is significantly worse than None in 8 out of 15 settings (with medium or large effect size). When comparing the median values of accuracy, FairBalance is only slightly worse than None (1-2%). This amount of decrease in Accuracy is as expected after applying a bias mitigation algorithm [7, 17, 21].
- In every setting, the fairness metrics (AOD, EOD, and SPD) for every sensitive attributes are significantly reduced after applying FairBalance. This also indicates that FairBalance is compatible with all the base classifiers under test.
- The computational overhead for FairBalance is around 1 second on compas dataset, 10 seconds on adult dataset, and <1 second on german dataset. Given the size of datasets in Table 3, this observation is consistent with the analysis that the computational cost for FairBalance is $O(n)$.

The above three observations suggest that:

Answer to RQ1: FairBalance can reduce bias on every sensitive attribute while maintaining similar prediction performances with very low computational overhead— $O(n)$ and good compatibility to different classifiers.

RQ2. Can FairBalance balance class distributions (FairBalanceClass) as well?

Comparing the performance of FairBalanceClass with None and FairBalance, we observe:

- F_1 score of FairBalanceClass is significantly higher than that of None and FairBalance in every setting (except for NB). Accuracy of FairBalanceClass is significantly lower than that of FairBalance, which is as expected. This suggests that class balancing in FairBalanceClass is effective in trading accuracy for higher F_1 score on the minority class.
- In every setting, the fairness metrics (AOD, EOD, and SPD) for every sensitive attributes are significantly reduced (compared to None) after applying FairBalanceClass.

The above two observations suggest that:

Answer to RQ2: FairBalanceClass can reduce bias on every sensitive attribute while increasing the prediction performance on the minority class (when data is imbalanced).

RQ3. How does FairBalance perform compared to the existing state-of-the-art bias mitigation algorithms?

As described in Section 4, some of the baseline bias mitigation algorithms only work with single sensitive attribute. We split the experimental results into two tables. Table 5 shows the results of FairBalance and other bias mitigation algorithms focusing on single sensitive attribute. From this table, we observe:

- Overall, FairBalance and FairBalanceClass are the best treatments according to the total ranks. This is because that the other treatments often perform poorly in terms of metrics on the non-target attribute. For example, Reweighing: race has very low AOD, EOD, and SPD on race but high AOD, EOD, and SPD on sex.
- Runtime of FairBalance and FairBalanceClass is orders of magnitude lower than the other treatments except for Reweighing. Runtime of Reweighing is a bit lower than is the proposed method but they are at the same order of magnitude ($O(n)$).
- The median results of each fairness metric from FairBalance and FairBalanceClass are all below 0.10 on all three datasets. This

Table 4: Comparisons of performances before and after FairBalance. Each dataset has two sensitive attributes. The second sensitive attribute is "Age" for the German dataset and "Race" for the other two. Medians (IQRs) are reported for 50 repeats. Colored cells represent whether they are significantly better than the chosen baseline (None) with effect size of small, medium, or large or significantly worse than the chosen baseline (None) with effect size of small, medium, or large.

Base Classifier	Dataset	Treatment	F_1	Accuracy	Runtime (sec)	Sex			Race / Age		
						AOD	EOD	SPD	AOD	EOD	SPD
LR	compas	None	60 (1)	67 (1)	1 (0)	16 (3)	20 (5)	19 (2)	16 (2)	20 (5)	17 (2)
		FairBalance	60 (1)	66 (1)	2 (0)	3 (4)	4 (7)	6 (4)	6 (7)	9 (8)	9 (7)
		FairBalanceClass	62 (1)	65 (0)	2 (0)	-2 (5)	-1 (5)	1 (5)	-2 (5)	0 (5)	0 (4)
	adult	None	48 (1)	80 (0)	10 (0)	-28 (1)	-45 (1)	-21 (0)	-11 (1)	-17 (3)	-10 (0)
		FairBalance	50 (0)	78 (0)	23 (0)	0 (1)	0 (3)	-6 (0)	0 (1)	1 (3)	-4 (1)
		FairBalanceClass	55 (0)	73 (1)	22 (1)	0 (4)	-1 (5)	-8 (4)	-2 (2)	-1 (5)	-6 (2)
	german	None	21 (6)	70 (3)	0 (0)	16 (9)	25 (14)	12 (8)	38 (19)	48 (17)	36 (20)
		FairBalance	2 (6)	70 (3)	0 (0)	0 (1)	0 (3)	0 (1)	0 (1)	0 (2)	0 (1)
		FairBalanceClass	50 (4)	59 (3)	0 (0)	2 (6)	4 (13)	4 (7)	6 (13)	4 (14)	11 (14)
SVM	compas	None	60 (1)	66 (1)	0 (0)	14 (4)	19 (6)	16 (3)	14 (3)	19 (4)	17 (2)
		FairBalance	60 (2)	66 (1)	2 (0)	4 (5)	4 (5)	8 (5)	7 (8)	9 (7)	10 (9)
		FairBalanceClass	62 (1)	65 (1)	2 (0)	0 (5)	0 (6)	3 (5)	-2 (4)	0 (4)	0 (4)
	adult	None	48 (0)	80 (0)	13 (1)	-27 (0)	-45 (0)	-20 (0)	-10 (1)	-16 (3)	-9 (0)
		FairBalance	49 (8)	79 (0)	24 (0)	-7 (8)	-12 (13)	-10 (6)	0 (4)	2 (6)	-3 (3)
		FairBalanceClass	54 (0)	73 (1)	24 (1)	0 (5)	0 (6)	-7 (5)	-1 (2)	0 (4)	-6 (2)
	german	None	16 (15)	70 (2)	0 (0)	12 (14)	20 (25)	11 (12)	26 (39)	34 (48)	24 (38)
		FairBalance	0 (3)	70 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)	0 (0)
		FairBalanceClass	49 (4)	60 (3)	0 (0)	4 (7)	5 (10)	4 (8)	7 (7)	3 (11)	12 (7)
DT	compas	None	60 (2)	66 (1)	0 (0)	14 (5)	17 (8)	17 (6)	15 (4)	19 (4)	17 (4)
		FairBalance	60 (1)	66 (1)	2 (0)	3 (7)	5 (9)	6 (6)	7 (6)	11 (6)	10 (5)
		FairBalanceClass	61 (1)	64 (1)	2 (0)	3 (8)	4 (7)	6 (8)	-4 (8)	-1 (10)	-2 (8)
	adult	None	49 (2)	80 (0)	8 (0)	-28 (2)	-46 (3)	-21 (2)	-10 (2)	-17 (4)	-10 (1)
		FairBalance	42 (1)	78 (0)	19 (0)	8 (2)	13 (3)	0 (0)	0 (4)	1 (7)	-2 (1)
		FairBalanceClass	53 (0)	71 (0)	19 (0)	1 (5)	0 (5)	-5 (4)	4 (5)	6 (6)	-1 (6)
	german	None	22 (6)	70 (2)	0 (0)	17 (7)	23 (11)	15 (8)	36 (6)	48 (8)	33 (6)
		FairBalance	18 (13)	68 (3)	0 (0)	9 (14)	10 (23)	9 (12)	17 (26)	22 (37)	16 (23)
		FairBalanceClass	48 (4)	58 (3)	0 (0)	7 (13)	10 (15)	9 (12)	-7 (19)	-5 (20)	-4 (22)
RF	compas	None	60 (2)	66 (1)	17 (0)	14 (7)	17 (9)	17 (6)	14 (4)	19 (4)	17 (5)
		FairBalance	60 (1)	66 (1)	19 (0)	4 (5)	6 (5)	7 (5)	6 (6)	10 (6)	9 (6)
		FairBalanceClass	61 (1)	64 (0)	19 (0)	3 (11)	4 (13)	6 (11)	-2 (7)	0 (7)	0 (7)
	adult	None	48 (1)	80 (0)	106 (1)	-28 (1)	-45 (1)	-21 (0)	-10 (1)	-16 (3)	-9 (0)
		FairBalance	42 (1)	78 (0)	118 (2)	8 (1)	13 (3)	0 (0)	0 (2)	2 (4)	-2 (1)
		FairBalanceClass	54 (0)	71 (1)	117 (1)	0 (5)	0 (5)	-6 (4)	2 (5)	4 (6)	-4 (4)
	german	None	25 (5)	70 (2)	10 (0)	18 (8)	27 (13)	16 (7)	36 (9)	49 (11)	33 (10)
		FairBalance	20 (11)	69 (2)	10 (0)	13 (17)	17 (26)	11 (16)	27 (23)	36 (37)	26 (22)
		FairBalanceClass	50 (3)	59 (3)	10 (0)	5 (10)	7 (10)	7 (10)	-5 (17)	-6 (15)	-3 (14)
NB	compas	None	64 (1)	66 (1)	0 (0)	35 (2)	39 (4)	38 (2)	23 (3)	27 (5)	25 (3)
		FairBalance	62 (1)	64 (1)	2 (0)	9 (2)	7 (4)	13 (2)	12 (2)	14 (2)	14 (2)
		FairBalanceClass	62 (1)	64 (1)	2 (0)	8 (2)	8 (4)	12 (3)	13 (3)	14 (4)	15 (3)
	adult	None	50 (0)	56 (1)	8 (0)	-6 (1)	-5 (1)	-14 (1)	-4 (2)	-3 (1)	-8 (3)
		FairBalance	48 (0)	53 (0)	18 (0)	0 (1)	0 (1)	-5 (1)	-1 (1)	-1 (1)	-5 (1)
		FairBalanceClass	45 (3)	46 (7)	18 (0)	2 (1)	1 (1)	-1 (3)	-1 (1)	-1 (1)	-4 (1)
	german	None	48 (5)	62 (3)	0 (0)	17 (14)	22 (18)	19 (11)	26 (16)	24 (18)	29 (14)
		FairBalance	47 (4)	61 (3)	0 (0)	7 (11)	9 (13)	8 (8)	12 (7)	12 (16)	15 (5)
		FairBalanceClass	51 (4)	60 (3)	0 (0)	4 (6)	5 (12)	5 (6)	8 (8)	7 (9)	12 (8)

suggests that applying either FairBalance or FairBalanceClass with logistic regression classifier can reduce the machine learning bias to a very low level.

The above observations show that it is very necessary to mitigate machine learning bias on multiple sensitive attributes simultaneously. For this reason, FairBalance is a better choice than the other baseline treatments on data with multiple sensitive attributes.

Next, we compare FairBalance against baseline treatments which are able to mitigate machine learning bias on multiple sensitive attributes simultaneously. From Table 6, we observe:

Table 5: Comparisons between FairBalance and baseline bias mitigation algorithms focusing on single sensitive attribute. Logistic regression classifier is utilized as the base classifier if the treatment does not specify a classifier. The second column in Treatment represent the target sensitive attribute. *Rank: Medians (IQRs)* are reported for 50 repeats (10 repeats for Reject Option Classification due to memory leak). Treatments of the same rank are not significantly different (p-value for Mann-Witney Utest is larger than 5% or effect size is smaller than medium). Lower rank the better. All numbers reported are in percentage. The Column Total Rank sums up rankings of all metrics except for Runtime. Colored cells represent top ranks in each metric R0 and R1 .

Dataset	Treatment		F_1	Accuracy	Sex			Race / Age			**Total Rank**	Runtime (sec)
					AOD	EOD	SPD	AOD	EOD	SPD		
compas	Reweighing	sex	R1: 60 (1)	R0: 66 (0)	R1: 7 (5)	R2: 11 (6)	R1: 9 (5)	R2: 14 (3)	R4: 19 (5)	R3: 17 (3)	14	R1: 1 (0)
		race	R1: 61 (1)	R0: 66 (0)	R3: 19 (4)	R3: 22 (5)	R3: 22 (3)	R0: 0 (10)	R1: 4 (11)	R0: 1 (10)	11	R0: 1 (0)
	Fair-SMOTE	sex	R0: 61 (2)	R1: 65 (1)	R1: -5 (7)	R1: -3 (7)	R0: -1 (7)	R0: 5 (10)	R2: 9 (10)	R1: 7 (10)	6	R5: 579 (82)
		race	R0: 62 (1)	R1: 65 (1)	R3: 20 (6)	R3: 21 (9)	R3: 23 (4)	R0: -6 (6)	R1: -2 (5)	R0: -3 (6)	11	R4: 395 (64)
	Adversial	sex	R1: 61 (2)	R0: 66 (1)	R5: 27 (11)	R5: 33 (15)	R5: 29 (9)	R3: 19 (5)	R5: 24 (7)	R4: 21 (5)	28	R3: 226 (20)
	Debiasing	race	R1: 60 (2)	R0: 66 (1)	R2: 14 (9)	R3: 19 (10)	R2: 17 (8)	R1: 11 (8)	R3: 15 (8)	R2: 14 (8)	14	R3: 221 (19)
	Reject Option	sex	R0: 62 (2)	R1: 66 (1)	R0: -2 (9)	R1: 2 (15)	R0: 0 (7)	R3: 19 (4)	R5: 22 (4)	R4: 21 (4)	14	R6: 1029 (66)
	Classification	race	R0: 62 (2)	R1: 66 (0)	R4: 22 (3)	R4: 23 (7)	R4: 25 (2)	R0: -1 (5)	R1: 2 (10)	R0: 0 (5)	14	R6: 1058 (43)
	FairBalance		R1: 60 (2)	R1: 66 (1)	R0: 4 (5)	R1: 5 (6)	R1: 8 (5)	R0: 3 (10)	R1: 6 (11)	R0: 6 (10)	5	R2: 2 (0)
	FairBalanceClass		R0: 62 (1)	R1: 65 (1)	R0: -1 (5)	R0: -1 (5)	R0: 2 (5)	R0: -3 (3)	R0: 0 (5)	R0: -1 (3)	1	R2: 2 (0)
adult	Reweighing	sex	R5: 49 (0)	R2: 78 (0)	R0: 0 (1)	R0: 0 (3)	R1: -6 (0)	R2: -14 (9)	R3: -21 (17)	R3: -12 (4)	16	R0: 14 (0)
		race	R6: 49 (1)	R0: 80 (0)	R2: -28 (0)	R2: -46 (0)	R4: -21 (0)	R1: 0 (1)	R0: 0 (3)	R0: -4 (0)	15	R1: 16 (1)
	Fair-SMOTE	sex	R2: 54 (0)	R6: 69 (2)	R0: -1 (1)	R0: -2 (1)	R3: -9 (1)	R3: -26 (4)	R4: -26 (5)	R5: -28 (5)	23	R7: 7710 (2871)
		race	R0: 57 (0)	R4: 72 (0)	R3: -32 (1)	R1: -35 (2)	R5: -37 (0)	R1: -2 (2)	R0: 0 (3)	R2: -8 (2)	16	R7: 8863 (1507)
	Adversial	sex	R4: 50 (0)	R1: 79 (0)	R1: -3 (2)	R0: -5 (4)	R3: -8 (1)	R2: -14 (3)	R3: -21 (4)	R3: -13 (2)	17	R4: 1445 (33)
	Debiasing	race	R6: 48 (1)	R0: 80 (0)	R2: -29 (0)	R2: -46 (1)	R4: -22 (0)	R1: 0 (4)	R0: 1 (6)	R0: -3 (3)	15	R4: 1445 (31)
	Reject Option	sex	R3: 53 (0)	R6: 70 (1)	R1: 4 (4)	R0: 3 (4)	R0: -4 (4)	R2: -17 (2)	R2: -17 (5)	R4: -21 (1)	18	R6: 5954 (242)
	Classification	race	R1: 57 (0)	R5: 71 (0)	R4: -34 (0)	R1: -37 (2)	R6: -39 (1)	R1: 2 (2)	R1: 3 (2)	R0: -3 (3)	19	R5: 5666 (52)
	FairBalance		R4: 50 (1)	R2: 78 (0)	R0: 0 (7)	R0: 0 (12)	R2: -6 (3)	R0: 0 (1)	R0: 1 (3)	R0: -4 (1)	8	R3: 19 (0)
	FairBalanceClass		R2: 54 (1)	R3: 73 (1)	R0: 0 (5)	R0: -2 (5)	R3: -8 (4)	R1: -1 (1)	R0: 0 (2)	R1: -6 (2)	10	R2: 18 (0)
german	Reweighing	sex	R3: 12 (12)	R0: 69 (3)	R1: 0 (11)	R1: 0 (14)	R1: 0 (10)	R2: 20 (24)	R2: 20 (29)	R3: 20 (23)	13	R0: 0 (0)
		age	R4: 4 (6)	R0: 69 (2)	R1: 1 (3)	R1: 2 (4)	R1: 1 (3)	R0: -1 (2)	R1: -1 (4)	R0: 0 (1)	8	R0: 0 (0)
	Fair-SMOTE	sex	R0: 50 (5)	R1: 62 (2)	R1: 2 (9)	R1: 0 (13)	R2: 4 (8)	R2: 14 (11)	R2: 13 (16)	R2: 18 (10)	11	R3: 103 (25)
		age	R1: 50 (5)	R1: 60 (3)	R1: 2 (8)	R1: 4 (12)	R2: 2 (10)	R1: 4 (12)	R1: 1 (15)	R1: 7 (13)	9	R4: 123 (24)
	Adversial	sex	R2: 28 (19)	R0: 68 (8)	R3: 13 (80)	R3: 22 (107)	R3: 11 (66)	R2: 21 (35)	R3: 26 (54)	R3: 20 (30)	19	R2: 67 (1)
	Debiasing	age	R2: 35 (14)	R1: 63 (34)	R2: -4 (38)	R2: -8 (49)	R2: -1 (37)	R3: -26 (144)	R4: -30 (148)	R3: -20 (143)	19	R2: 68 (0)
	Reject Option	sex	R0: 53 (4)	R1: 60 (2)	R1: 3 (6)	R1: 0 (10)	R2: 5 (6)	R2: 16 (14)	R2: 14 (16)	R3: 22 (9)	12	R5: 272 (10)
	Classification	age	R1: 49 (4)	R1: 58 (4)	R2: 6 (9)	R1: 5 (8)	R2: 8 (7)	R1: 1 (9)	R1: -1 (11)	R1: 5 (12)	10	R5: 268 (3)
	FairBalance		R4: 4 (7)	R0: 70 (2)	R0: 0 (1)	R0: 0 (2)	R0: 0 (1)	R0: 0 (2)	R0: 0 (5)	R0: 0 (1)	4	R1: 0 (0)
	FairBalanceClass		R1: 50 (4)	R1: 59 (3)	R1: 1 (7)	R1: 0 (10)	R2: 2 (7)	R1: 5 (8)	R1: 4 (13)	R1: 10 (9)	9	R1: 0 (0)

- On the german dataset, only **Fair-SMOTE-Multiple** and **FairBalanceClass**, the two treatment with class balancing, achieved acceptable F_1 scores. The other treatments all have close to 0 F_1 scores on the minority class. This suggests that the learned model predicted (almost) every data point as the majority class. Therefore, although **Fair-SMOTE-Multiple** and **FairBalanceClass** have the worst total ranks, they are the best and the only acceptable treatments on the german dataset. This also suggests that class balancing is very important on datasets with severe class imbalance.
- On compas and german datasets, **FairBalanceClass** performed similarly with **Fair-SMOTE-Multiple**. However, **FairBalanceClass** outperformed **Fair-SMOTE-Multiple** in the adult dataset.
- **FairBalanceClass** outperformed **FERMI** on all three datasets.
- **FairBalanceClass** outperformed **Reweighing-Multiple** on compas and german datasets. On the adult datasets, these two treatments achieved the same total ranks.
- **FairBalance** outperformed **Reweighing-Multiple** on compas and adult datasets. On the german datasets, these two treatments achieved the same total ranks.
- **FairBalanceClass** outperformed **FairBalance** on compas and german datasets while **FairBalance** outperformed **FairBalanceClass** on the adult dataset.
- **FairBalance** and **FairBalanceClass** have the shortest runtime on every dataset.

From the above observations, **FairBalanceClass** outperformed all other treatments except for **FairBalance**. Overall, we would recommend using **FairBalanceClass** to avoid critical failures such as the result of **FairBalance** on the german dataset.

Answer to RQ3: FairBalanceClass outperformed the existing state-of-the-art bias mitigation algorithms. It is recommended to apply FairBalanceClass to mitigate bias and balance the classes on datasets with multiple sensitive attributes and unbiased labels.

Table 6: Comparisons between FairBalance and baseline bias mitigation algorithms focusing on multiple sensitive attributes. FERMI: 30K represents its hyperparameter $\lambda = 30,000$ while FERMI: 10K represents $\lambda = 10,000$. *Rank: Medians (IQRs)* are reported for 50 repeats (10 repeats for Reject Option Classification due to memory leak). Treatments of the same rank are not significantly different (p-value for Mann-Witney Utest is larger than 5% or effect size is smaller than medium). Lower rank the better. All numbers reported are in percentage. The Column Total Rank sums up rankings of all metrics except for Runtime. Colored cells represent top ranks in each metric R0 and R1. Red cells highlight the critical failures on the german dataset.

Dataset	Treatment	F_1	Accuracy	Sex			Race / Age			**Total Rank**	Runtime (sec)
				AOD	EOD	SPD	AOD	EOD	SPD		
compas	Fair-SMOTE-Multiple	R0: 62 (1)	R1: 65 (1)	R0: 0 (10)	R0: 0 (8)	R0: 2 (9)	R0: -2 (5)	R0: 1 (7)	R0: 0 (5)	1	R3: 400 (54)
	FERMI: 30K	R0: 62 (1)	R1: 65 (0)	R0: 1 (3)	R0: 2 (6)	R1: 4 (3)	R1: -6 (3)	R0: -3 (4)	R0: -4 (4)	3	R2: 53 (5)
	FERMI: 10K	R0: 62 (1)	R1: 65 (1)	R0: 4 (4)	R0: 4 (7)	R2: 7 (4)	R0: -4 (5)	R0: -1 (6)	R0: -2 (4)	3	R2: 51 (2)
	Reweighing-Multiple	R1: 60 (1)	R0: 66 (1)	R0: 2 (3)	R0: 3 (5)	R2: 7 (3)	R1: 5 (4)	R1: 9 (6)	R1: 8 (4)	6	R1: 2 (0)
	FairBalance	R1: 60 (2)	R0: 66 (1)	R0: 4 (5)	R0: 5 (6)	R2: 8 (5)	R0: 3 (10)	R1: 6 (11)	R1: 6 (10)	5	R0: 2 (0)
	FairBalanceClass	R0: 62 (1)	R1: 65 (1)	R0: -1 (5)	R0: -1 (5)	R0: 2 (5)	R0: -3 (3)	R0: 0 (5)	R0: -1 (3)	1	R0: 2 (0)
adult	Fair-SMOTE-Multiple	R0: 55 (0)	R2: 72 (1)	R2: -9 (6)	R1: -10 (6)	R3: -17 (6)	R2: -8 (6)	R2: -8 (7)	R2: -13 (5)	14	R5: 6002 (83)
	FERMI: 30K	R2: 51 (0)	R4: 64 (1)	R1: 3 (4)	R0: 3 (4)	R0: -3 (4)	R2: 7 (4)	R2: 8 (4)	R0: 3 (4)	11	R4: 1264 (23)
	FERMI: 10K	R0: 55 (0)	R3: 69 (0)	R3: -26 (1)	R2: -26 (2)	R4: -33 (1)	R2: -9 (3)	R1: -6 (3)	R2: -15 (2)	17	R3: 63 (0)
	Reweighing-Multiple	R3: 50 (7)	R0: 78 (0)	R0: 0 (5)	R0: -2 (9)	R1: -6 (1)	R1: 1 (4)	R1: 4 (7)	R0: -3 (3)	6	R2: 24 (1)
	FairBalance	R3: 50 (1)	R0: 78 (0)	R0: 0 (7)	R0: 0 (12)	R1: -6 (3)	R0: 0 (1)	R0: 1 (3)	R0: -4 (1)	4	R1: 19 (0)
	FairBalanceClass	R1: 54 (1)	R1: 73 (1)	R0: 0 (5)	R0: -2 (5)	R2: -8 (4)	R1: -1 (1)	R0: 0 (2)	R1: -6 (2)	6	R0: 18 (0)
german	Fair-SMOTE-Multiple	R0: 49 (5)	R1: 61 (3)	R2: 3 (7)	R2: 6 (10)	R2: 5 (10)	R2: 9 (13)	R2: 10 (18)	R2: 11 (13)	13	R4: 121 (5)
	FERMI: 30K	R2: 0 (0)	R0: 69 (2)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	2	R3: 32 (0)
	FERMI: 10K	R2: 0 (0)	R0: 69 (3)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	R0: 0 (0)	2	R2: 31 (0)
	Reweighing-Multiple	R1: 2 (5)	R0: 69 (2)	R1: 0 (1)	R1: 0 (0)	R1: 0 (0)	R1: 0 (1)	R1: 0 (3)	R1: 0 (1)	7	R1: 0 (0)
	FairBalance	R1: 4 (7)	R0: 70 (2)	R1: 0 (1)	R1: 0 (2)	R1: 0 (1)	R1: 0 (2)	R1: 0 (5)	R1: 0 (1)	7	R0: 0 (0)
	FairBalanceClass	R0: 50 (4)	R1: 59 (3)	R2: 1 (7)	R2: 0 (10)	R2: 2 (7)	R2: 5 (8)	R2: 4 (13)	R2: 10 (9)	13	R0: 0 (0)

RQ4. Are (1) difference in group size and (2) difference in class distribution within each group the main reasons for machine learning bias when class labels are unbiased?

(1) The fact that **FairBalance** outperformed **Reweighing-Multiple** in Table 6 indicates that difference in group size is a reason causing the machine learning bias when class labels are unbiased.
(2) The fact that the fairness metrics of **Reweighing-Multiple** in Table 6 are much lower than those of **None** with logistic regression classifier in Table 4 indicates that difference in class distribution within each group is another reason causing the machine learning bias when class labels are unbiased.
(3) Difference in class distribution within each group is more important for machine learning bias than difference in group size. This is because the reduction of fairness metrics after balancing the class distributions within each group is much larger than after balancing the group sizes.

Answer to RQ4: Yes, both (1) difference in group size and (2) difference in class distribution within each group are reasons for machine learning bias when class labels are unbiased. Difference in class distribution within each group is a more important reason than difference in group size.

6 DISCUSSION

6.1 Limitations

The scope of this work limits its application. The results and conclusions of this work are limited to

- Datasets with unbiased class labels.
- Known sensitive attributes.

6.2 Threats to validity

Sampling Bias - We have used three datasets in this work. We could find only these three datasets that have more than one sensitive attribute. We have experimented with five different classification models. Conclusions may change a bit if other datasets and models are used.

Evaluation Bias - We used the three most popular fairness metrics in this study. Prior works [10, 16, 19] only used one or two metrics although IBM AIF360 contains more than 50 metrics. In future, we will explore more evaluation criteria.

Conclusion Validity - One assumption of evaluating our experiments is that the test data is unbiased. Prior fairness studies also made similar assumption [3, 7]. This assumption is true for the three datasets we used since their class labels were derived from truth. On other datasets with human decisions as class labels, this assumption may fail and we will need other ways to make sure the test data is unbiased.

External Validity - This work focuses on classification problems which are very common in AI software. We are currently working on extending it to regression models. The scope of this work also limits it generalizability to problems with unknown sensitive attributes and potentially biased class labels.

7 CONCLUSION

This paper proposes FairBalance, a pre-processing technique with low computational overhead for mitigating machine learning bias on multiple sensitive attributes. Our results show that, within the scope of this paper, FairBalance can significantly reduce machine learning bias while maintaining the prediction performance. Also,

its class balancing variant FairBalanceClass consistently outperforms other state-of-the-art bias mitigation algorithms on datasets with multiple sensitive attributes. Note that the scope of this paper also limits the usage of FairBalance and FairBalanceClass—it cannot reduce the bias inherited from biased training labels. Our results also validated the hypothesis that on datasets with unbiased ground truth labels, ethical biases in the learned models largely attribute to the training data having (1) difference in group size and (2) difference in class distribution within each group.

To sum up, the best practice suggested in this paper is to apply FairBalance when the dataset has unbiased labels and multiple known sensitive attributes, and to apply FairBalanceClass if the classes imbalanced as well.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[3] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Nov 2020). https://doi.org/10.1145/3368089.3409704

[4] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[5] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3995–4004.

[6] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*. 319–328.

[7] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in Machine Learning Software: Why? How? What to Do?. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Athens, Greece) *(ESEC/FSE 2021)*. Association for Computing Machinery, New York, NY, USA, 429–440. https://doi.org/10.1145/3468264.3468537

[8] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: a way to build fair ML software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 654–665.

[9] Joymallya Chakraborty, Kewen Peng, and Tim Menzies. 2020. Making Fair ML Software Using Trustworthy Explanation. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering* (Virtual Event, Australia) *(ASE '20)*. Association for Computing Machinery, New York, NY, USA, 1229–1233. https://doi.org/10.1145/3324884.3418932

[10] Joymallya Chakraborty, Tianpei Xia, Fahmid M. Fahid, and Tim Menzies. 2019. Software Engineering for Fairness: A Case Study with Hyperparameter Optimization. arXiv:1905.05786 [cs.SE]

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[12] Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin* 114, 3 (1993), 494.

[13] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[16] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs.LG]

[17] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[18] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting Reject Option in Classification for Social Discrimination Control. *Inf. Sci.* (2018). https://doi.org/10.1016/j.ins.2017.09.064

[19] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 35–50.

[20] Arjun Kharpal. 2018. Health care start-up says A.I. can diagnose patients better than humans can, doctors call that 'dubious'. https://www.cnbc.com/2018/06/28/babylon-claims-its-ai-can-diagnose-patients-better-than-doctors.html.

[21] Andrew Lowy, Rakesh Pavan, Sina Baharlouei, Meisam Razaviyayn, and Ahmad Beirami. 2021. FERMI: Fair Empirical Risk Minimization via Exponential Renyi Mutual Information. *arXiv preprint arXiv:2102.12586* (2021).

[22] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[23] Parmy Olson. 2011. The Algorithm That Beats Your Bank Manager. https://www.forbes.com/sites/parmyolson/2011/03/15/the-algorithm-that-beats-your-bank-manager/#15da2651ae99.

[24] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. arXiv:1709.02012 [cs.LG]

[25] Andrew Jhon Scott and M Knott. 1974. A cluster analysis method for grouping means in the analysis of variance. *Biometrics* (1974), 507–512.

[26] E. Strickland. 2016. Doc bot preps for the O.R. *IEEE Spectrum* 53, 6 (June 2016), 32–60. https://doi.org/10.1109/MSPEC.2016.7473150

[27] Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1715–1724.

[28] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[29] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.